A Systematic Review of Computational Approaches to Deciphering Bronze Age Aegean and Cypriot Scripts

Maja Braović
University of Split / FESB
Department of Electronics
and Computing
maja.braovic@fesb.hr

Damir Krstinić
University of Split / FESB
Department of Electronics
and Computing
damir.krstinic@fesb.hr

Maja Štula
University of Split / FESB
Department of Electronics
and Computing
maja.stula@fesb.hr

Antonia Ivanda
University of Split / FESB
Department of Electronics
and Computing
antonia.senta.00@fesb.hr

This article provides a detailed insight into computational approaches for deciphering Bronze Age Aegean and Cypriot scripts, namely, the Archanes script and the Archanes formula, Phaistos Disk, Cretan hieroglyphic (including the Malia Altar Stone and Arkalochori Axe), Linear A, Linear B, Cypro-Minoan, and Cypriot scripts. The unique contributions of this article are threefold: (1) a thorough review of major Bronze Age Aegean and Cypriot scripts and inscriptions, digital data and corpora associated with them, existing computational decipherment methods developed in order to decipher them, and possible links to other scripts and languages; (2) the definition of

Action Editor: Xuanjing Huang. Submission received: 11 September 2023; revised version received: 16 January 2024; accepted for publication: 25 February 2024.

https://doi.org/10.1162/coli_a_00514

15 major challenges that can be encountered in computational decipherments of ancient scripts; and (3) an outline of a computational model that could possibly be used to simulate traditional decipherment processes of ancient scripts based on palaeography and epigraphy. In the context of this article the term decipherment denotes the process of discovery of the language and/or the set of symbols behind an unknown script, and the meaning behind it.

1. Introduction

The process of determination of the accurate time frame of when writing first appeared is extremely difficult. It is currently believed that writing first appeared in 3200 BCE in the region of Sumer (Glassner 2018), located in southern ancient Mesopotamia (present-day Iraq). The script that appeared in this region is known today as Sumerian pre-cuneiform, and this writing served as a foundation for the later Sumerian cuneiform. After the appearance of Sumerian cuneiform other scripts emerged in Egypt (approximately 3100 BCE), Indus Valley (approximately 2500 BCE, although it is arguable whether this was a linguistic script or not [Sproat 2014; Oakes 2019]) in present-day India and Pakistan, Crete (approximately 1900 BCE) in present-day Greece, China (approximately 1200 BCE), and Central America (approximately 600 BCE) (Robinson 2007). The first sign of writing west of Egypt is usually considered to be the Archanes script discovered in the 1960s in the Aegean (Decorte 2018). Not much is known about this script, and it is not generally agreed upon whether it is indeed a script or a repeated collection of symbols known as the Archanes formula. Regardless of its nature, the Archanes script and the Archanes formula were precursors to Europe's most famous scripts used during the Bronze Age (c. 3000-700 BCE [Vandkilde 2016]), and possibly some of the earliest European scripts—Cretan hieroglyphic, Phaistos Disk, Linear A, Linear B, Cypro-Minoan, and Cypriot scripts. Many decipherment attempts of these scripts have been proposed over the years, but only the Linear B and Cypriot scripts have generally been accepted as deciphered. Decipherment attempts of ancient scripts have usually been based on epigraphy (the study of ancient inscriptions), paleography (the study of ancient handwritings), grammatical comparison to possibly related scripts, and so forth, and only recently have the computational methods aimed at the decipherment of these scripts started appearing.

Even though the term **decipherment** usually refers to the process of determination of the symbol system behind an unknown script or text, in the context of this paper it will denote the process of discovery of the language and/or the set of symbols behind an unknown script. According to Gelb and Whiting (1975), decipherments in their broader sense can be classified into four types that are based on the level of the expert knowledge regarding the writing system and the language involved: (0) both the language and the writing system are known, (I) the language is known but the writing system is unknown, (II) the language is unknown but the writing system is known, and (III) both the language and the writing system are unknown. In Gelb and Whiting (1975) it is emphasized that the "known" and "unknown" concepts in this context are not solid, but rather fuzzy categories that shade into each other. The first of these types of decipherment (Type 0) that Gelb and Whiting (1975) suggested is trivial and does not represent a problem at all, the second (Type I) and the third (Type II) are difficult, and the fourth type (Type III) is the most difficult one of them all and might even be undecipherable. By taking this categorization into account, we can categorize Bronze Age Aegean and Cypriot scripts and languages. Linear B and Cypriot syllabaries would fall into the Type 0 category, since the symbol sets and the languages behind

them are already known. Linear A syllabary would fall into the Type II category, since the language behind it is unknown, but the set of symbols is mostly known as it is related to Linear B. The Archanes script and the Archanes formula, Phaistos disk, Cretan hieroglyphic script, and the Cypro-Minoan syllabary would all unfortunately fall into the Type III category, since neither the language nor the majority of their symbols are currently known. Attempts at computational decipherment of these scripts should focus on discovering the language(s) behind them, a complete set of symbols, and the writing systems they were used in.

This article presents a systematic review of the Archanes script and the Archanes formula, Phaistos Disk, Cretan hieroglyphic, Linear A, Linear B, Cypro-Minoan, and Cypriot scripts, and focuses on the computational approaches proposed for their decipherment. Even though Linear B and Cypriot scripts are already considered to be deciphered, computational approaches to their decipherment still exist because they might prove useful to the decipherments of other scripts possibly related to them but still undeciphered. In addition, we give an overview of available digital corpora related to the Bronze Age Aegean and Cypriot scripts and discuss challenges one might face when attempting to decipher them. As far as we are aware, this is the first systematic review that focuses only on the techniques proposed for the computational decipherment of Bronze Age Aegean and Cypriot scripts. The only similar review we were able to find was presented in Ferrara and Tamburini (2022), and although it is an interesting resource that focuses on paleographic, epigraphic, and computational techniques for the decipherment of ancient scripts, as well as the challenges that the researchers might face, it does not discuss many of the existing methods for the computational decipherment of Bronze Age Aegean and Cypriot scripts, nor the digital corpora or the related languages that might prove useful in the decipherment process.

We would like to fully disclose that this article is not written from the standard epigraphical or paleographical point of view. Its primary target audience comprises computer science researchers for whom archaeology, linguistics, epigraphy, or any of the related fields mainly associated with the traditional methods for the decipherment of ancient scripts are not primary research areas. This article is primarily intended for researchers interested in using computational techniques to decipher ancient scripts, as there has been a surge in the use of natural language processing (NLP) and deep learning techniques in the last few years. It is our belief that these novel resources available to computer scientists could prove useful in the decipherment of ancient scripts. Besides these main goals, this article can also serve as an introduction to some of the earliest European scripts and languages.

The article is structured as follows. In Section 2 we discuss the human language, linguistic classification of language families, and various types of classifications regarding the different types of writing systems. In Section 3 we discuss Bronze Age Aegean and Cypriot scripts, namely, the Archanes script and the Archanes formula, Cretan hieroglyphic (including the Malia Altar Stone and Arkalochori Axe), Phaistos Disk, Linear A, Linear B, Cypro-Minoan, and Cypriot scripts. In Section 4 we discuss challenges that need to be addressed when attempting to decipher ancient, unknown scripts. In Section 5 we discuss digital resources and corpora available for the Bronze Age Aegean and Cypriot scripts. In Section 6 we give a detailed overview of traditional and deep learning based approaches commonly used in NLP. In Section 7 we give an overview of computational approaches to the decipherment of Bronze Age Aegean and Cypriot scripts, alongside an overview of computational approaches proposed for the decipherment of other scripts as well. Additionally, we look at possible methods that can be used to simulate traditional processes used for the decipherment of ancient

scripts. In Section 8 we discuss languages and scripts that were geographically close and contemporary with the Bronze Age Aegean and Cypriot scripts, with hopes that some of them can be indisputably linked to the undeciphered Bronze Age Aegean and Cypriot scripts and aid in their decipherment. In Section 9 we give a conclusion and discuss future work.

2. Languages and Writing Systems

Human language, or the capacity for it, is probably at least 150,000 to 200,000 years old (Pagel 2017). Approximately 7,000 languages exist in the world today (Radikov et al. 2019), and most of them are classified as belonging to certain language families. Language families represent groups of languages that share some similarities and have the same ancestral language, usually called a proto-language. For example, English language (alongside most European languages) belongs to the Indo-European language family, whose ancestral language is hypothesized to be some unknown Proto-Indo-European[1] language. Even though most of the existing languages can be classified as belonging to one of the known language families, there are languages that are distinct and unique enough to be classified as language isolates because their connection to known language families cannot be reliably established (e.g., Basque and Burushaski languages [Urban 2021]). For a detailed overview of the world's languages and language families we refer the reader to Pereltsvaig (2021), and for a list of language families and the languages they encompass we refer the reader to Glottolog (2023).

As means for writing down various languages, many different writing systems and scripts have been developed over the years. A *writing system* is a notational system for a natural language (Neef 2015), while its physical representation is referred to as a *script* (Pae and Wang 2022). Writing systems can be divided into various categories based on the phonetics or semantics connected with the symbols that they are associated with. For example, a writing system where symbols represent vowels and consonants can be referred to as a type of alphabet called a *phonemic alphabet*, whose specific instances include, for example, Latin and Greek scripts. A writing system where symbols represent syllables instead of vowels or consonants can be referred to as a *syllabary*, and an example of a script belonging to this type of a writing system is Cherokee (Cushman 2011).

Even though we mentioned phonemic alphabets and syllabaries, we need to emphasize that there is no universally agreed upon taxonomy of writing systems, and many researchers in this field propose different classifications. For example, Robinson (2009) divided writing systems into 6 different categories: syllabic systems (e.g., Japanese Kana), logosyllabic systems (e.g., Chinese), logoconsonantal systems (e.g., Egyptian), consonantal alphabets (e.g., Arabic), phonemic alphabets (e.g., Greek), and logophonemic alphabets (e.g., English). Allan (2015) divided writing systems into logographic and phonographic scripts, and split them into even smaller subcategories based on the size of phonetic units that their symbols represent. Sproat and Gutkin (2021) divided writing systems into syllabic systems (e.g., Chinese), moraic systems (e.g., Japanese Kana), consonantal systems or abjads (e.g., Semitic scripts), abugidas or alphasyllabaries (e.g., Brahmic scripts), and alphabets (e.g., Greek). Gelb and Whiting (1975) divided full writing systems into logo-syllabic, syllabic, and alphabetic. Daniels

1 More information on Proto-Indo-European language can be found in Rau (2010).

and Bright (2010) divided writing systems into logosyllabaries, syllabaries, consonatories or abjads, alphabets, and abugidas.

It is important to note that many scripts cannot fully be classified as belonging to just one type of a writing system, regardless of its name. Many scripts can contain properties that are characteristic of multiple types of writing systems, and this is why there is no universally agreed upon taxonomy of writing systems. In this article, however, we will use the taxonomy of writing systems presented in Daniels and Bright (2010), which divides writing systems into logosyllabaries, syllabaries, consonatories or abjads, alphabets, and abugidas. **Logosyllabaries** are types of writing systems where "the characters of a script denote individual words (or morphemes) as well as particular syllables" (Daniels and Bright 2010). Examples of logosyllabaries include the Mayan logosyllabary, cuneiform (Gnanadesikan 2009), Egyptian hieroglyphs, and Chinese. **Syllabaries** are types of writing systems where "the characters denote particular syllables, and there is no systematic graphic similarity between the characters for phonetically similar syllables" (Daniels and Bright 2010). Examples of syllabaries include Cretan hieroglyphic, Linear A, Linear B, Cypro-Minoan, Cypriot (Karnava 2014b), and Cherokee (Cushman 2011). **Abjads** or **consonatories** are types of writing systems where "the characters denote consonants (only)" (Daniels and Bright 2010). They are also referred to as consonantal alphabets since they do have symbols for consonants like true alphabets do, but, unlike true alphabets, not for the vowels. In abjads, the vowels are inferred either from the word's syntax or semantics, and not by explicitly stating them. Examples of abjads include Hebrew, Ugaritic, Mandaic, Estrangelo, and Arabic (Daniels 2003). **Alphabets** are types of writing systems where "the characters denote consonants and vowels" (Daniels and Bright 2010). Examples of alphabets include Latin (or Roman), Greek, Glagolitic, Cyrillic, Etruscan, and Phoenician. The Latin alphabet is the most widely used script in the world (Kolinsky and Verhaeghe 2017). **Abugidas** are types of writing systems where "each character denotes a consonant accompanied by a specific vowel, and the other vowels are denoted by a consistent modification of the consonant symbols" (Daniels and Bright 2010). They are also known as alphasyllabaries, since they are somewhere in between true alphabets and true syllabaries. Examples of abugidas include Thai, Burmese, Khmer, and Lao (Ding, Utiyama, and Sumita 2018).

In order to decipher an unknown script, it is important to first determine the type of writing system it belongs to. Based on characteristics of undeciphered scripts it is possible, with a certain degree of confidence, to determine the writing system they belong to. If, for example, a script only has a small number of symbols, it is probably an alphabet. Alphabets use symbols that represent either consonants or vowels, and therefore do not need a large set of symbols to represent words. However, if a script has a very large number of symbols, it is possible that they represent either syllables or ideas, and therefore that script may belong to a logosyllabary or a syllabary. According to Coe (1999) as cited in Fuls (2015), alphabets do not have more than 36 signs, while pure syllabaries have between 40 and 90. Scripts that have hundreds or thousands of signs are more complex and combine a small number of phonetic signs with a large number of logograms (Robinson 2009).

In this article we will focus on computational decipherment of Bronze Age Aegean and Cypriot scripts, usually classified as syllabaries. Potential decipherment of different types of writing systems would represent an entirely different problem than the one we are focused on and would require completely different viewpoints and methods to potentially realize. This falls out of scope of our current research, but would be an interesting area to focus on in the future.

3. Bronze Age Aegean and Cypriot Scripts and Inscriptions

Many ancient scripts and inscriptions still await decipherment, for example, the inscriptions written in the Etruscan alphabet (present-day Italy), Cretan hieroglyphic script (present-day Greece), Linear A (present-day Greece), Proto-Elamite script (present-day Iran), Rongorongo (present-day Easter Island), Indus script (present-day Pakistan and India), Vinča or Danube script (present-day Serbia and Romania), Sitovo inscription (present-day Bulgaria), and so forth. Deciphering these scripts is an extremely complicated process because researchers can encounter many obstacles in their way. These obstacles are outlined in more detail in Section 4, but can include things such as multiple unknown languages behind an unknown script, unknown writing systems, and unknown reading directions.

The decipherment of any of the world's undeciphered scripts and inscriptions would represent an enormous breakthrough. However, in this article, we are particularly focused on scripts and inscriptions found mostly in the Aegean area and in Cyprus.

Major Bronze Age Aegean and Cypriot scripts include the Archanes script and the Archanes formula, Cretan hieroglyphic script, Phaistos Disk, Linear A syllabary, Linear B syllabary, Cypro-Minoan syllabary, and the Cypriot syllabary. There are a few inscriptions and early signs of writing from the Aegean area that are not generally accepted as standalone scripts, and these include the Malia Altar Stone and Arkalochori Axe. Figure 1 shows the locations of the islands of Crete and Cyprus, where the majority of Bronze Age Aegean and Cypriot inscriptions were discovered. For the creation of the map we used QGIS geospatial software (QGIS Development Team 2023).

Figure 2 shows the timeline of Bronze Age Aegean and Cypriot scripts. The approximate dates of use for each of these scripts are taken from corresponding sources cited in Sections 3.1–3.7.

In subsections below we give an overview of Bronze Age Aegean and Cypriot scripts; for an extensive and in-depth review of Aegean pre-alphabetic scripts we refer the reader to Davis (2010).

3.1 The Archanes Script and the Archanes Formula

The earliest recognized evidence for the use of a script on Crete was found on a number of seals, and these inscriptions are known today as the Archanes formula (Anastasiadou 2016). Archanes formula refers to two sign groups, or sign sequences, present in their entirety or in part on a number of seals belonging to the wider collection of inscribed objects called the Archanes script (Olivier, Godart, and Poursat 1996; Karnava 1999; Decorte 2018). The objects belonging to the Archanes script were excavated by Yannis Sakellarakis in the mid 1960s near Archanes on Crete, and are generally dated somewhere between the end of the third and the beginning of the second millennium BCE (c. 2200–1800 BCE) (Decorte 2018). According to Ferrara and Weingarten (2022), the Archanes formula cannot be considered a proper writing system or a script. Out of all the Aegean scripts or inscriptions that have been analyzed, the Archanes script and the Archanes formula were the ones with the least number of attempts at their decipherment, possibly due to the small number of inscriptions.

3.2 Cretan Hieroglyphic Script

The Cretan hieroglyphic script, or Cretan hieroglyphs, dates back to the Middle Bronze Age in Crete (c. 2100/2050–1700/1675 BCE) (Nosch and Ulanowska 2021). Because this

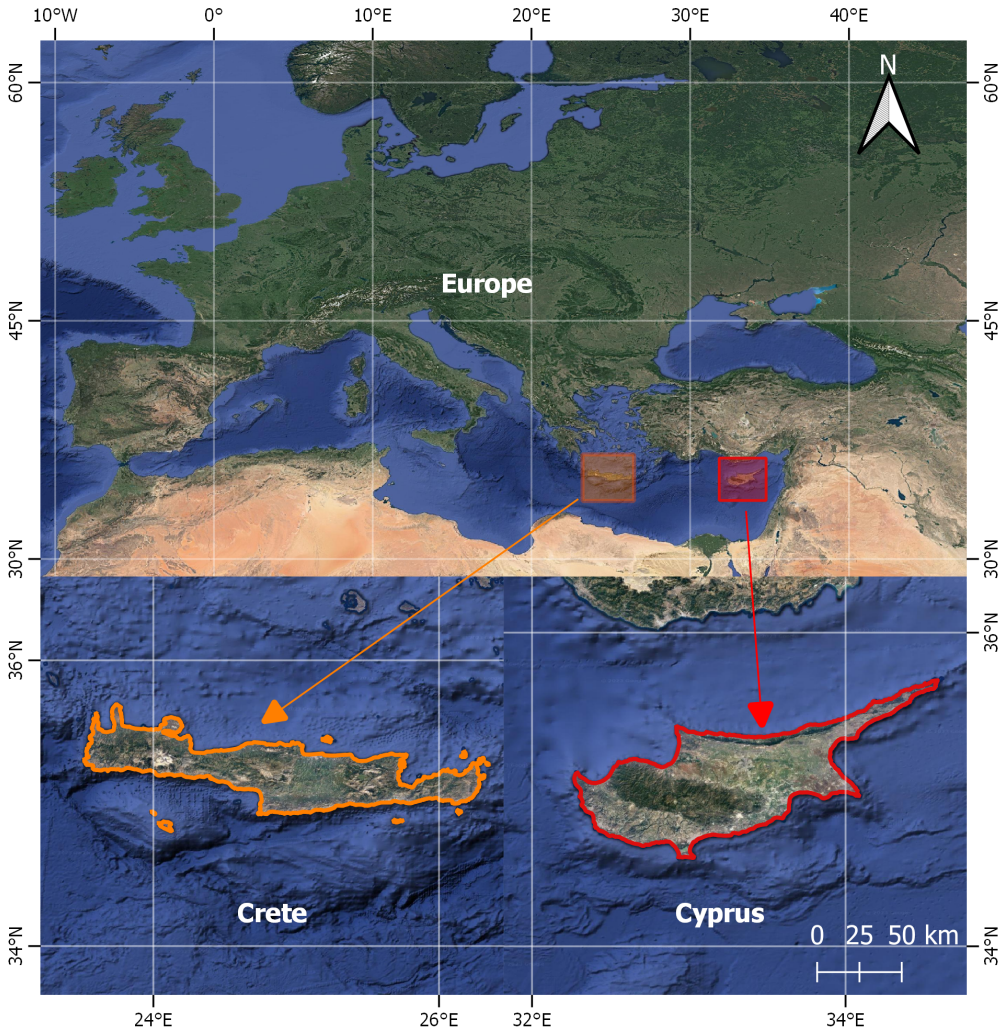


Figure 1
Crete and Cyprus are places where the majority of Bronze Age Aegean and Cypriot inscriptions were discovered.

script does not appear to be connected to other writing systems that used hieroglyphs in their inscriptions (e.g., Egyptian or Hittite), it is thought to have been invented locally on Crete (Karnava 2014a) rather than developed from other writing systems that predated them. This represents a significant obstacle in their potential decipherment—if they are not related to other known scripts or inscriptions, it might be incredibly difficult or even impossible to ever understand them. To add to the difficulties regarding the potential decipherment of this script, there are currently fewer than 400 known inscriptions that bear the Cretan hieroglyphs, and the inscriptions themselves are rather short (Ferrara, Montecchi, and Valério 2021).

It is generally assumed that the Cretan hieroglyphic script represents syllabic writing (Revesz 2022; Karnava 2014a) with a logographic element (Karnava 2014a), so it

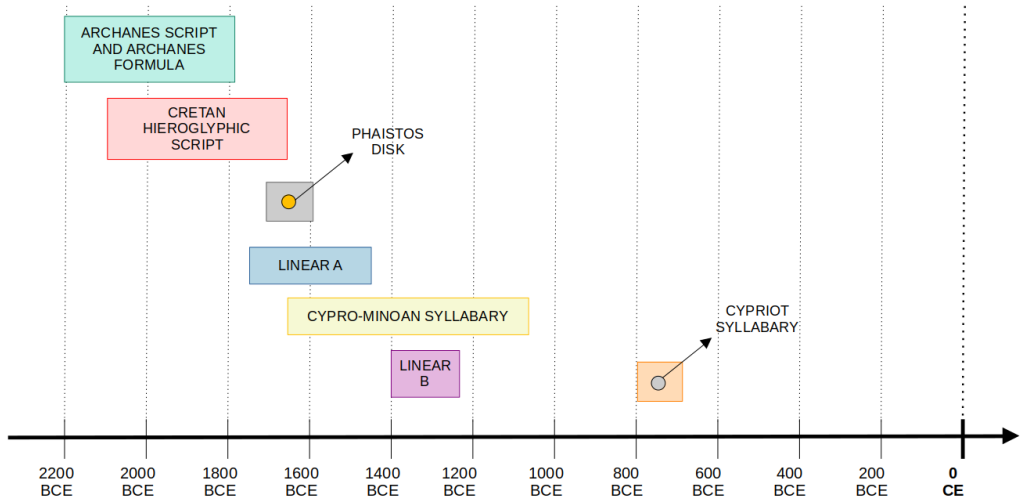


Figure 2
Approximate timeline of Bronze Age Aegean and Cypriot scripts.

can be classified as a logosyllabary (Civitillo 2021). According to Karnava (2014a), both right-to-left and left-to-right reading and writing directions are used in the Cretan hieroglyphic script. The language behind the Cretan hieroglyphic script is called a Minoan language (Revesz 2017b). This language was probably spoken by the Bronze Age inhabitants of Crete, and is currently still unknown.

Over the years, researchers have attempted to determine if the Cretan hieroglyphic script is related to other writing systems, inscriptions, or languages. They have examined various formulas and scripts, such as the Archanes formula (Ferrara, Montecchi, and Valério 2021), Linear A (Ferrara, Montecchi, and Valério 2022), Linear B (Daggumati and Revesz 2019), Proto-Hungarian and Hungarian languages (Revesz 2016b), and so on. In a study by Serafimov and Tomezzoli (2011) the authors lay evidence of early Slavic presence in Crete in the second millennium BCE, and note similarities between the Cretan hieroglyphic script, Linear A, and Linear B scripts with the Vinča (in present-day Serbia), Gradeshnitsa, and Karanovo (both in present-day Bulgaria) culture writings. Revesz (2016a) suggests that Crete is the likely origin of the Cretan Script Family, which includes the Cretan hieroglyphic script, Linear A, Linear B, Cypriot syllabary, Greek alphabet, Phoenician alphabet, Old Hungarian alphabet, South Arabic alphabet, and the Tifinagh alphabet. This script family was expanded in Revesz (2017a) with the addition of the Carian alphabet. Despite numerous theories about the origin of Cretan hieroglyphs and their connection to other scripts and languages, there is no single generally accepted theory, and Cretan hieroglyphic script remains a mystery.

Cretan hieroglyphic script and Linear A co-existed for almost two centuries (Ferrara, Montecchi, and Valério 2022), even though Cretan hieroglyphic script predates Linear A, and Linear A is based on it (Davis 2010). Cretan hieroglyphic script consists of 96 syllabograms, ten of which also serve as logograms (Nosch and Ulanowska 2021). This would make the Cretan hieroglyphic script logo-syllabic. A list of Cretan hieroglyphic signs can be found in Ferrara, Montecchi, and Valério (2023).

Some of the longest known inscriptions in Cretan hieroglyphs are the Phaistos Disk, the Malia Altar Stone, and the Arkalochori Axe (Revesz 2017c), but it is important to

emphasize that sometimes the Phaistos Disk and the Arkalochori Axe are listed as separate scripts (Revesz 2018). In the scope of this article we will regard the Phaistos Disk as a separate script, as this seems to be the current consensus, and Malia Altar Stone and Arkalochori Axe as texts written in Cretan Hieroglyphic script. The Malia Altar Stone (or Malia Stone Inscription) was excavated in 1937 in Malia on Crete, and it is the only known inscription in Cretan hieroglyphs engraved in stone (Kenanidis and Papakitsos 2017) (most of the other inscriptions are found on clay objects or sealstones [Younger 1999]). The Malia Altar Stone consists of 16 symbols (Revesz 2017d), and their translation was proposed in Revesz (2017c,d). The Arkalochori Axe is a bronze double axe discovered in 1935 in a cave in central Crete (Duhoux 1998). It contains 15 signs and dates back from c. 1700 BCE (Davis 2010). A translation of the Arkalochori Axe was proposed in Revesz (2017c).

The Cretan hieroglyphs are still generally regarded as undeciphered, although there have been attempts at their decipherment (e.g., Revesz 2016b).

3.3 Phaistos Disk

The Phaistos Disk is a unique, one-of-a-kind circular clay object discovered in the ruins of the Phaistos palace in 1908 by Luigi Pernier (Reczko 2009). It differs from other Aegean and Cypriot Bronze Age artefacts in terms of its construction, as it was imprinted with small seals rather than inscribed by hand. The diameter of the Phaistos Disk varies from 158 to 165 millimeters as it is not perfectly round, and its thickness varies from 16 to 21 millimeters (Evans 1909b). The disk dates back to the seventeenth century BCE (Chorozoglou, Koukis, and Papakitsos 2017) and is imprinted on both sides with 241 symbols, with a total of 45 unique symbols (Reczko 2009). The Phaistos Disk is still generally regarded as undeciphered, although there have been a few attempts at its decipherment (e.g., Achterberg 2004; Revesz 2015b, 2016c). The summary of decipherment attempts can be found in Eisenberg (2008). The reading direction of the disk has been debated and is still not generally agreed upon. Some of the signs imprinted on the disk have a stroke underneath and might represent diacritical marks, similar to the strokes used in Devanagari script for writing Sanskrit (Chadwick 1987). Although there is a general view among archaeologists that the Phaistos Disk is an authentic Bronze Age artefact, there have been times when this view was questioned and when it was suggested that the disk itself might be a forgery (Eisenberg 2008). The Phaistos Disk is currently on display in the Heraklion Archaeological Museum on Crete.

3.4 Linear A Script

Linear A is an undeciphered script that was discovered by the British archaeologist Sir Arthur J. Evans during his excavations in Greece in the late 1800s and early 1900s. It was used during the period c. 1750–1450 BCE (Robinson 2009), and is classified as a logosyllabary (Salgarella and Castellan 2021a; Salgarella 2022). This script is mainly associated with Crete, but was also used on other islands in the Aegean Sea as well as on Mainland Greece. The total number of signs from all 1,370 known Linear A documents is estimated to be between 7,362 and 7,396 (Schoep 2002). Out of all these signs, 97 are unique, and 64 of them are implemented in Linear B with slight modifications (Tan 2022). Average word length in Linear A documents is 3.3 signs (Fuls 2015). The majority of Linear A clay tablets come from the Minoan palace of Haghia Triada (Chadwick 1987).

The language behind Linear A is considered to be Minoan language, the same one as the one behind the Cretan hieroglyphic script (Revesz 2017b). A link between these two scripts in terms of their common origin was suggested in Owens (1996).

Since the language behind Linear A is still unknown, there have been many attempts to discover what this language is or where it originated from. Some of these attempts include trying to connect Linear A with Hurro-Urartian languages, the Indo-European language family, and the Etruscan language (Facchetti 2002). Duhoux (1990) states the following: "Linear A's language has been recognised as cognate to an impressive list of languages: Hittite, Luwian, Lycian, Sanskrit, Greek, Indo-European, Semitic, Carian, Basque, and so on." Recently, Revesz (2016c, 2020) proposed a theory that Linear A is connected to the Uralic language family. According to Schoep (2002), the best-founded hypotheses for the language behind Linear A are Semitic (non-Indo-European language) and Lycian (an extinct language belonging to the Anatolian branch of the Indo-European language family). We will discuss languages possibly related to Bronze Age Aegean and Cypriot scripts in more detail in Section 8.

3.5 Linear B Script

Linear B was discovered in the form of clay tablets and vessels by the British archaeologist Sir Arthur J. Evans during his excavations in Greece in the late 1800s and early 1900s. It was in use on Crete and Mainland Greece during the period c. 1400–1190 BCE (Salgarella 2020). It is classified as a logosyllabary (Salgarella and Castellan 2021a).

Linear B was deciphered in the 1950s by Michael Ventris and John Chadwick, with significant contributions from Emmett L. Bennett and Alice E. Kober (Bennett, Chadwick, and Ventris 1956; Chadwick and Ventris 1973; Ventris and Chadwick 2015; Kober 1948). It was used for writing Mycenaean Greek (Davis 2010).

Linear A and Linear B scripts have approximately the same number of signs (Melena 2014). Out of 97 unique Linear A signs, 64 of them are implemented in Linear B with slight modifications (Tan 2022). Syllabograms common to both Linear A and Linear B, alongside their phonetic transcriptions, are listed in Melena (2014). Although there is a clear link between Linear A and Linear B scripts, they were almost certainly used to write different languages (Whittaker 2005). Linear B was used for writing the Mycenaean Greek (Davis 2010) language, while the language written in Linear A script is still unknown but referred to as Minoan. Additionally, both scripts were primarily used for administrative purposes, but the usage of Linear A was wider than the one of Linear B, which seems to have been restricted to palatial bureaucracy (Whittaker 2005). For the detailed comparison of Linear A versus Linear B inscriptions in terms of places of discovery and objects on which the inscriptions themselves were found, the reader is encouraged to consult Tomas (2010).

3.6 Cypro-Minoan Syllabary

The Cypro-Minoan syllabary was the script used by the pre-Greek inhabitants of Cyprus (Davis 2010). Although its name implies that it was used only on Cyprus, it was also used in coastal parts of Syria and Tiryns in Greece (Valério 2016). The script was in use between 1550 and 1050 BCE (Smith and Hirschfeld 1999), and comes in three varieties that are linked to the places where the inscriptions were discovered (Billigmeier 1976). Not all researchers agree on the number of varieties of this script, and research presented in Corazza et al. (2022) indicates that it is a unitary script. The Cypro-Minoan syllabary is based on Linear A, has around 100 unique symbols, and is

still undeciphered (Davis 2011). Nearly 250 inscriptions in Cypro-Minoan are known to exist, and the total number of syllabograms than can be found in these inscriptions is less than 4,000 (Valério 2016).

3.7 Cypriot Syllabary

The Cypriot syllabary (or Classical Cypriot script) was mainly used on Cyprus between the eighth and third centuries BCE (Karnava 2014c). It was developed from the Cypro-Minoan syllabary by the people who lived on Cyprus and spoke their own dialect of the Greek language (Davis 2010). The syllabary consisted of either 54 signs (common syllabary) or 55 signs (Paphian syllabary) (Karnava 2014c). There are around 880 inscriptions bearing Cypriot syllabary writing, 800 of which were found on Cyprus and 80 in Egypt (Davis 2010). The Cypriot syllabary was deciphered, and its decipherment began in 1871 with the work of Assyriologist George Smith (Davis 2010), but later contributed to by many other researchers as well (Karnava 2014c). Cypriot syllabary was used to write the Arcado-Cypriot Greek language (Kenanidis and Papakitsos 2015).

4. Challenges in Deciphering Unknown Scripts

Luo et al. (2021) described the two main challenges that can be encountered during attempts to decipher lost languages: "(1) the scripts are not fully segmented into words; (2) the closest known language is not determined." While this is true, we would also like to expand this list with other challenges that should be addressed when attempting to decipher lost scripts, and propose the following:

1. *Unknown writing system and reading direction.* This obstacle is one of the greatest challenges in deciphering ancient scripts. Current practices commonly use vocabulary size to determine the type of writing system to which the script belongs. If this size is uncertain or unknown, there are no certain ways to determine whether the writing system is an alphabet, abugida, abjad, syllabary, logosyllabary, or a mixture of different writing systems. If a writing system is unknown, establishing the reading direction can also be difficult. There are many types of reading directions that scripts can have. For example, some scripts are read from right to left, while others are read from left to right. Some scripts use boustrophedon[2] style of writing (e.g. the Great Inscription on Crete that is part of the Gortyn law code [Remondino et al. 2008]). Sometimes, the reading direction can be inferred from scripts that are geographically and chronologically close to the script being deciphered. However, this method does not always work, as some scripts are autochthonous and not related to other scripts.

2 Boustrophedon writing is defined in Merriam-Webster.com Dictionary (2023a) as "the writing of alternate lines in opposite directions (as from left to right and from right to left)." When given enough input data, this kind of writing is relatively simple to recognize as it uses systematic mirror imaging or flipping of individual glyphs.

2. *Unknown punctuation.* Not knowing whether the script uses punctuation or spaces between the words can make the undeciphered text difficult to segment into words and subsequently decipher.
3. *Unknown language.* Not being familiar with the language or languages the script is used to encode can make the script almost undecipherable. This is unfortunately the case with most of the currently undeciphered scripts such as the Cretan hieroglyphic script, Linear A, and Cypro-Minoan script.
4. *Small dataset size.* When a dataset of available inscriptions in undeciphered script is very small it might be impossible to decipher that script. It should be remembered that the archaeologists are still discovering ancient artefacts and that there is hope the small datasets of undeciphered inscriptions will be expanded in the near future. Recently, in 2022, scientists succeeded in translating the oldest sentence written in the world's first alphabet (Vainstub et al. 2022; Osborne 2022), and this gives us hope that new discoveries, translations, and decipherments of ancient inscriptions will follow in the future.
5. *Incomplete vocabulary.* Sometimes the vocabulary of an unknown script is incomplete, usually due to the small number of inscriptions or texts available. This makes it difficult to determine the type of writing system the script possibly belongs to, and complicates the potential frequency analysis that can be performed on the text in order to gain more information about its content, language, or grammar.
6. *Unknown syntactic (or word order) typology.* Not being familiar with the dominant order of different types of words (e.g., verbs, subjects, and objects) in a sentence can pose a significant problem in a decipherment process. In linguistics, the word order in a sentence can be classified into six traditional typologies regarding the relative position of subject (S), object (O), and verb (V): SVO, SOV, VSO, VOS, OVS, and OSV (Croft 2003). However, these are not the only existing word order typologies as others have been proposed as well. For example, Dryer (1997) proposed a word order typology that classifies languages as OV or VO, and as SV or VS. Additionally, free word order typologies also exist (e.g., Slavic languages are generally classified as free SVO languages [Siewierska and Uhlířová 1998]), and can further complicate the decipherment process of unknown scripts.
7. *Unknown morphological (or word structure) typology.* Not knowing how or if the words in undeciphered scripts can change their form can aggravate the decipherment process. In morphological typology there are two distinct types of languages, analytic and synthetic (Šincek 2020), that different languages can be classified into. Synthetic languages, which are more flexible than the analytic ones, can further be divided into four subtypes: agglutinating, fusional or inflected, polysynthetic, and oligosynthetic languages (Šincek 2020). Agglutinating languages, which combine smaller morphemes into words, are particularly interesting as they include many ancient languages (e.g., Sumerian, Elamite, Hattic [Tóth 2007]) that are possibly related to undeciphered Aegean and Cypriot scripts.

8. *Existence of allographs.* Not knowing whether there are multiple symbols or signs used for the same sound, syllable, word, concept, or idea can aggravate the decipherment process. For example, in the English alphabet there are uppercase and lowercase letters for each sound, but if someone unfamiliar with that tried to decipher the text written in the English language, they would certainly encounter a number of difficulties in their decipherment process. Their constructed chart of possible symbols used in the undeciphered text would be twice as long as it realistically needs to be, and that would probably lead to many incorrect conclusions. In this example we did not even consider different writing styles such as cursive, which would certainly further complicate the entire process.

9. *Difficulties regarding the penmanship of each particular scribe.* Another problem that is closely related to the previous one regards the handwriting or penmanship of each particular scribe that wrote or inscribed ancient texts. Depending on their handwriting, the same symbols could look very different, and may lead to further difficulties in constructing the complete set of symbols connected to a certain script.

10. *Existence of exonyms.* Not knowing whether exonyms exist in the text can make the decipherment process of undeciphered text more difficult. Exonyms are usually localized names for foreign places that differ from their autochthonous names. For example, *Croatia* is an exonym (a name for Croatia used by the English-speaking people), while *Hrvatska* is an endonym (a name for Croatia used by the Croatian people).

11. *Existence of homonyms.* Not knowing whether the words in undeciphered scripts change their meaning based on the context they are used in can have an impact on the decipherment process. These words are known as homonyms, and can be either spelled or pronounced the same. Since it is usually unknown how to accurately pronounce words that belong to undecipherd scripts, in this context we consider the homonyms to represent words that differ in meaning and have the same spelling, while their pronunciation can be considered unknown.

12. *Unknown context.* Not knowing the exact location where the undeciphered texts were discovered can have an adverse effect on the decipherment process. The location where the texts are found can provide significant contextual information about the content of the text. For example, if the inscribed clay tablet with unknown text is found inside an old church, it is far more likely to contain religious text than bureaucratic text.

13. *Unavailable parallel data.* In the context of NLP, parallel data usually encompasses the same text written in different languages or scripts. The lack of this data makes the decipherment process very difficult, and is unfortunately the reality for all of the undeciphered scripts today as far as we are aware. For example, Egyptian hieroglyphs presented a major decipherment problem and were only successfully deciphered when the Rosetta Stone bearing the parallel inscriptions in three different languages was unearthed.

14. *Unavailable digital data and corpora.* The majority of undeciphered languages and scripts lack an adequate associated digital dataset that could be used in their potential computational decipherment.

15. *Unavailable hardware resources.* Computational decipherment of unknown ancient texts is an extremely challenging task, and it may be impossible to accomplish on limited hardware. The decipherment process could therefore necessitate the use of high performance computing (HPC), which is not readily available to everyone.

From the above outlined challenges we can conclude that many different things can affect the decipherment process of unknown scripts. The more information is known about the unknown script and the people that used it and their environment, the more likely it is that it will one day be deciphered.

The above outlined challenges in the decipherment processes of unknown scripts can be encountered whether the unknown script is being deciphered by either the computational or standard approaches based on epigraphy and paleography. One challenge, however, can only be associated with the computational approaches to unknown script decipherment, and that challenge is related to the computational power that is available for that task. Building flexible computational decipherment frameworks for automated decipherment of undeciphered scripts these days is almost inseparable from powerful hardware resources, both computationally and in terms of data storage. HPC solutions provide adequate computation platform. Scaling deciphering models, no matter what kind of approach is used to build a model (e.g., deep learning neural networks, machine learning methods, statistical models) is facilitated with the HPC platforms. When additional parameters are needed to extend possibilities and accuracy of the existing models or when the newly discovered texts for decipherment should be added to the existing ones, hardware resources that easily support such additions and extensions are a good choice. HPC resources today are much more accessible for research and for business applications, technically as well as financially, either as on-site solutions or provided by different public institutions and finally through private vendors with cloud computing resources like Microsoft Azure (Microsoft 2023), Amazon Web Services (Amazon 2023), Google Cloud Platform (Google 2023), and many others.

For additional information on helpful guidelines on strategies that can help in script decipherment processes the reader can refer to Valério (2016) and Ferrara and Tamburini (2022).

5. Data and Corpora

Traditional or printed corpora of Bronze Age inscriptions related to the Minoan civilization first started appearing in the early 1900s, with works such as Evans (1909a,b), and continued with a number of other corpora such as GORILA (Godart and Olivier 1985) and CHIC (Olivier, Godart, and Poursat 1996). Out of all the available traditional corpora related to the Bronze Age Minoan civilization, GORILA and CHIC corpora are the ones that are most frequently used. The GORILA corpus obtained its name from the first letters of surnames of its authors (**G**odart and **O**livier), and from the first letters of the first part of the title of their manuscript (**R**ecueil des **I**nscriptions en **L**inéaire **A**). The CHIC corpus obtained its name from the first letters of words included in the title of the manuscript where it was presented (**C**orpus **H**ieroglyphicarum **I**nscriptionum **C**retae).

Many of the digital corpora related to the Bronze Age Minoan civilization is based on the CHIC or GORILA corpora. The most well-known available digital corpora on ancient Bronze Age Aegean and Cypriot scripts (Aurora 2015a; Revesz, Rashid, and Tuyishime 2019b; Rashid 2019; Salgarella and Castellan 2021a; Ferrara et al. 2023a; Lastilla, Ravanelli, and Ferrara 2019; Younger 2023; Petrolito et al. 2015; Papavassileiou, Owens, and Kosmopoulos 2020; Greco, Flouda, and Notti 2023; Hogan 2022, 2023) are discussed below in more detail.

Aurora (2015a) presented DĀMOS (**Da**tabase of **M**ycenaean at **Os**lo), an electronic corpus of Mycenaean texts written in Linear B script. Alongside Mycenaean texts, this database includes information about scribal hands, sites where the texts were discovered, approximate time frames, phonological transcriptions of the texts, and so forth. The author plans on enriching the database with closely related digital resources, for example, the ones that contain information about Minoan Linear A. The DĀMOS database is currently available at Aurora (2015b).

Revesz, Rashid, and Tuyishime (2019b) and Rashid (2019) presented the AIDA (**A**ncient **I**nscription **D**atabase and **A**nalytics) system, which currently stores inscriptions written on the Phaistos Disk and the ones written in Linear A and Cretan hieroglyphic scripts. The authors plan on expanding the AIDA database with inscriptions written in Sumerian, Elamite, and the Indus Valley script. AIDA is currently available at Revesz, Rashid, and Tuyishime (2019a), where it provides possible syllabic values, translations, cognates, and related languages of Linear A symbols and sequences.

Salgarella and Castellan (2021a) presented SigLA (**Sig**ns of **L**inear **A**) online database of Linear A symbols and sequences that includes their graphical representations (drawings), phonetic transcriptions, places of origin, time period, types of artefacts on which the inscriptions are found, and so forth. The SigLA dataset is available at Salgarella and Castellan (2021b).

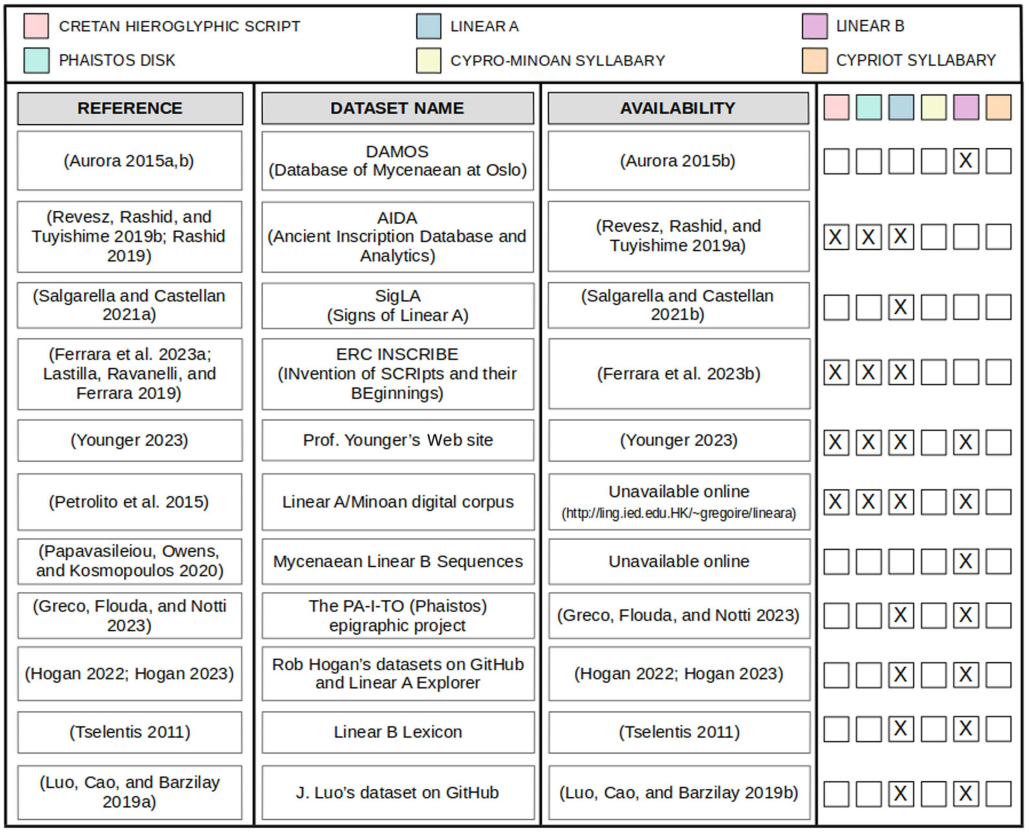
Through the ERC INSCRIBE (**IN**vention of **SCRI**pts and their **BE**ginnings) project (Ferrara et al. 2023a; Lastilla, Ravanelli, and Ferrara 2019), many 3D representations of artefacts bearing Cretan hieroglyphic and Linear A inscriptions are available online at Ferrara et al. (2023b). The ERC INSCRIBE project is still ongoing and it is possible that the online resources available through it will be expanded in the future.

Younger (2023) is a Web site created and maintained by professor emeritus John G. Younger. This Web site contains links to many online resources (some of which are written by Prof. Younger) regarding the Phaistos Disk, Linear A, Linear B, and Cretan hieroglyphic scripts. Some of the linked resources contain, among much valuable information, information on possible phonetic transcriptions of Linear A sequences.

Other digital corpora has been presented as well, for example, Petrolito et al. (2015) (Linear A), and Papavassileiou, Owens, and Kosmopoulos (2020) (Linear B), but unfortunately we were unable to find these datasets online. Additional online resources regarding the Bronze Age Aegean scripts include Greco, Flouda, and Notti (2023) (Linear A and Linear B), Tselentis (2011) (Linear B), Luo, Cao, and Barzilaya (2019a,b) (Linear B; this dataset is a modification of the one presented in Tselentis [2011]), Hogan (2022) (Linear A and Linear B), and a tool for exploring Linear A syllabary presented in Hogan (2023) and associated with Hogan (2022).

Figure 3 shows an overview of available online resources regarding the Bronze Age Aegean and Cypriot scripts.

It can be seen from the discussion above that there is a significant lack of digital resources associated with Bronze Age Aegean and Cypriot scripts, and this represents a significant challenge. It is also important to note that even though some of the digital resources regarding the Bronze Age Aegean and Cypriot scripts do exist, they are by no



REFERENCE	DATASET NAME	AVAILABILITY	
(Aurora 2015a,b)	DAMOS (Database of Mycenaean at Oslo)	(Aurora 2015b)	☐☐☐☐☒☐
(Revesz, Rashid, and Tuyishime 2019b; Rashid 2019)	AIDA (Ancient Inscription Database and Analytics)	(Revesz, Rashid, and Tuyishime 2019a)	☒☒☒☐☐☐
(Salgarella and Castellan 2021a)	SigLA (Signs of Linear A)	(Salgarella and Castellan 2021b)	☐☐☒☐☐☐
(Ferrara et al. 2023a; Lastilla, Ravanelli, and Ferrara 2019)	ERC INSCRIBE (INvention of SCRIpts and their BEginnings)	(Ferrara et al. 2023b)	☒☒☒☐☐☐
(Younger 2023)	Prof. Younger's Web site	(Younger 2023)	☒☒☒☐☒☐
(Petrolito et al. 2015)	Linear A/Minoan digital corpus	Unavailable online (http://ling.ied.edu.HK/~gregoire/lineara)	☒☒☒☐☒☐
(Papavasileiou, Owens, and Kosmopoulos 2020)	Mycenaean Linear B Sequences	Unavailable online	☐☐☐☐☒☐
(Greco, Flouda, and Notti 2023)	The PA-I-TO (Phaistos) epigraphic project	(Greco, Flouda, and Notti 2023)	☐☐☒☐☒☐
(Hogan 2022; Hogan 2023)	Rob Hogan's datasets on GitHub and Linear A Explorer	(Hogan 2022; Hogan 2023)	☐☐☒☐☒☐
(Tselentis 2011)	Linear B Lexicon	(Tselentis 2011)	☐☐☒☐☒☐
(Luo, Cao, and Barzilay 2019a)	J. Luo's dataset on GitHub	(Luo, Cao, and Barzilay 2019b)	☐☐☒☐☒☐

Figure 3
An overview of available online resources regarding the Bronze Age Aegean and Cypriot scripts.

means standardized or complete. It is an unfortunate fact that we cannot be completely certain of the completion of a set of symbols belonging to a particular script, because often only scarce or short inscriptions written in that script are available to us. One example of this is the Phaistos Disk, which is a unique object that cannot be definitely linked to anything else. Another problem related to the standardization of the set of symbols belonging to a particular undeciphered ancient script is that many of the symbols in digital databases are hand drawn by computer scientists. This is a direct consequence of the fact that digital representations of ancient symbols are usually not freely available and can be copyrighted. This represents a certain obstacle in constructing digitalized datasets that could be compatible with each other, as each available dataset will usually have its own set of hand-drawn symbols or contain phonemic transcriptions of symbols.

6. Commonly Used Methods in Natural Language Processing

Commonly used methods in NLP and subsequently in the automatic decipherment of ancient scripts can be divided into three categories: methods commonly associated with pre-processing of textual information, methods commonly associated with processing of textual information, and evaluation techniques designed to output a numerical

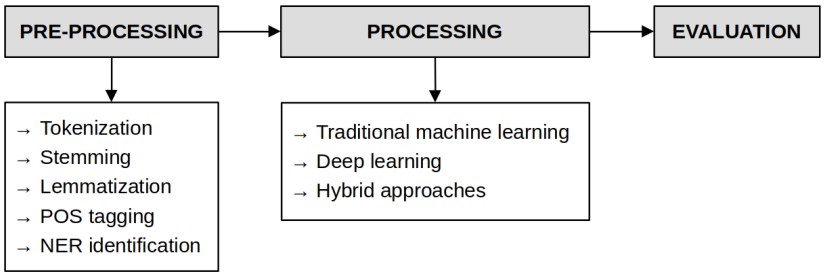


Figure 4
Flow diagram of common NLP methods.

representation that indicates how well the final NLP model is performing on novel textual data it has not seen during the training phase. Figure 4 shows a flow diagram of common NLP methods discussed in the following sections.

6.1 Pre-processing of Textual Information

Pre-processing of textual information is of utmost importance in NLP applications. Methods commonly associated with this step are used to prepare textual data for further processing, which subsequently helps in improving the accuracy and speed of the model. Pre-processing techniques in NLP encompass statistic and linguistic text analysis, and commonly include processes such as tokenization, stemming, lemmatization, POS (Part-of-Speech) tagging, and NER (Named Entity Recognition) identification.

6.1.1 Tokenization. Tokenization can be defined as the deconstruction of words into smaller tokens such as individual symbols or syllables (e.g., the word "disk" could be split into "di" and "sk"). Tokenization is a general term that encompasses sentence tokenization (the process of splitting the input text into sentences), word tokenization (the process of splitting the input text into words), subword tokenization (the process of splitting the input text into subwords), and character-based tokenization (the process of splitting the input text into singular characters such as letters, spaces and punctuation marks). Even though each of these tokenization methods can be used as a pre-processing step in NLP, by far the most widely used one is subword tokenization.

Subword tokenization has some advantages over the other tokenization methods in terms of vocabulary size and out-of-vocabulary words. In NLP, vocabulary can be defined as the number of unique tokens (words, subwords, etc.) that the model learns during the training phase, and subsequently uses in further processing. Subword tokenization keeps the vocabulary size relatively stable, while sentence and word tokenization methods tend to have much larger vocabularies, which increases the overall complexity of the model and decreases its velocity. Character-based tokenization methods have the smallest vocabularies, but also fail to deliver any contextual information to the model, which hinders its performance. Subword tokenization also deals very well with situations in which it encounters out-of-vocabulary words. These words are words that may appear in the input text given to the model, but that have not been encountered by the model during the training. In cases like these, sentence and word tokenization methods would fail, but character-level tokenization would not (given that the newly

encountered word does not contain any novel characters), and subword tokenization most likely would not fail either (although this cannot be guaranteed given its limited vocabulary size and the subword tokenization method used).

Most widely used subword tokenization methods include Byte-Pair Encoding (BPE) (Gage 1994; Sennrich, Haddow, and Birch 2016), WordPiece (Schuster and Nakajima 2012; Wu et al. 2016), and Unigram (Kudo 2018) models. The BPE model was initially designed for data compression purposes (Gage 1994), and later adapted for POS tagging (Sennrich, Haddow, and Birch 2016). It works by finding characters and/or groups of characters that most commonly appear together in the training dataset, and merges them into unique tokens that are used for vocabulary construction. The BPE model is most notably used in OpenAI's GPT-2 (Das and Verma 2020). The WordPiece model was initially constructed for Japanese and Korean voice search system (Schuster and Nakajima 2012), and later adapted for POS tagging (Wu et al. 2016). According to Wu et al. (2016), the WordPiece model is similar to Sennrich, Haddow, and Birch (2016), and is "data-driven to maximize the language-model likelihood of the training data, given an evolving word definition." The Unigram POS tagger, on the other hand, is a probabilistic "mixture of characters, subwords and word segmentations" (Kudo 2018), and is based on the Unigram language model. For more in-depth information on tokenization methods in NLP we would encourage the reader to consult Spathis and Kawsar (2023) and Mielke et al. (2021).

6.1.2 Stemming. Stemming can be defined as the process of determination of the stem of the word (i.e., the base of the word to which affixes such as prefixes and suffixes can be added; e.g. the stem of the word "decoding" is "decod"). Stemming is incredibly important when processing highly inflected languages (e.g., Croatian, Slovenian, Polish, and Finnish) because it keeps the vocabulary size under control by recognizing different forms of the word as variations of the same word instead of an entirely new word.

Stemming algorithms (or stemmers) can in general be classified as being either linguistic-based or computational-based (Jabbar et al. 2020). Linguistic-based stemmers use hand-crafted grammatical rules to determine the stem of the word (e.g., Porter 1980; Kariyawasam, Senanayake, and Haddela 2019; Kaur and Buttar 2020), while computational-based stemmers use statistical (i.e., machine learning-based; e.g., Melucci and Orio 2003; Bölücü and Can 2019) or non-statistical (i.e., corpus-based; e.g., Singh and Gupta 2017, 2019) computations (Jabbar et al. 2020).

It is important to note that even though many different stemming algorithms exist, not all of them will work well for all languages, as languages vary and can be highly specific and unique.

6.1.3 Lemmatization. Lemmatization can be defined as the process of determination of the lemma of the word (i.e., its canonical or dictionary form; e.g., the lemma of the word "decoding" is "decode"). It is similar to stemming, but inherently more complex and time-consuming. In terms of accuracy, it is difficult to determine which of the two is better, as studies have shown inconsistent results when comparing the two (Pramana et al. 2022).

Lemmatization algorithms (or lemmatizers) are usually either rule-based (e.g., Plisson et al. 2004; Stanković et al. 2016; Nandathilaka, Ahangama, and Weerasuriya 2018), machine learning-based (e.g., Kestemont et al. 2017; Freihat et al. 2018; Manjavacas, Kadar, and Kestemont 2019; Akhmetov et al. 2020; Karwatowski and

Pietron 2022; Hafeez et al. 2023) or they use a combination of these methods and represent hybrid models (e.g., Ingason et al. 2008; Sahala et al. 2023). Even though many traditional machine learning techniques have been used for lemmatization throughout the years (e.g., random forest classification model [Akhmetov et al. 2020] and maximum entropy classifier [Freihat et al. 2018]), it seems that the researchers have lately been more oriented towards the deep learning-based approaches (e.g., Kestemont et al. 2017; Manjavacas, Kadar, and Kestemont 2019; Ezhilarasi and Maheswari 2021b; Karwatowski and Pietron 2022; Hafeez et al. 2023).

Lemmatizers have even been developed for ancient inscriptions such as those written in Ancient Greek (Vatri and McGillivray 2020), Early Irish (i.e., Old and Middle Irish) (Dereza 2018), Classical Armenian, Old Georgian and Syriac (Vidal-Gorène and Kindt 2020), Akkadian (Sahala et al. 2023), and additionally for palaeographic eleventh century stone inscriptions as well (Ezhilarasi and Maheswari 2021b).

6.1.4 POS Tagging. POS tagging or grammatical tagging can be defined as the process of determination of parts of speech (nouns, verbs, adjectives, etc.) for words appearing in textual documents that are being examined. POS tagging is usually accomplished by using rule-based (or linguistic) approaches, stochastic (or probabilistic) approaches, or hybrid approaches.

Rule-based approaches to POS tagging generally encompass pre-defined grammatical rules determined by linguistic experts, and they usually focus on one specific language. This is to ensure that the rules determined for one language are not used for the POS tagging of another, as they are often not compatible. If a large enough textual dataset tagged with POS labels exists, then a simple algorithm can be constructed that could automatically behave (more or less successfully) like a human expert and construct grammatical rules useful for POS tagging. One of the simplest examples of this kind of an approach is Brill's tagger (Brill 1992). Brill's tagger is a simple yet powerful algorithm that extracts grammatical rules from a tagged corpus, and uses these rules in order to accomplish POS tagging of novel textual documents. Even though this tagger is usually classified as belonging to rule-based approaches to POS tagging, it is sometimes classified as belonging to machine learning based approaches to POS tagging (Chiche and Yitagesu 2022), which are inherently stochastic.

Stochastic approaches to POS tagging are used for determining the probabilities with which a word, given a certain context, belongs to each of the predefined POS tags. These approaches encompass commonly used machine learning algorithms such as artificial neural networks (e.g., Vidal-Gorène and Kindt 2020; Ezhilarasi and Maheswari 2021a,b), Hidden Markov Models (e.g., Lee, Tsujii, and Rim 2000; Gao and Johnson 2008; Stratos, Collins, and Hsu 2016), Support Vector Machines (e.g., Nakagawa, Kudo, and Matsumoto 2001; Giménez and Màrquez 2004; Ekbal and Bandyopadhyay 2008), and Conditional Random Fields (e.g., Krishnapriya et al. 2014). Artificial neural networks are mathematical models designed to mimic a human brain. They generally require a large amount of training data and computational resources, but often produce relatively accurate results as well. Hidden Markov Models (HMMs) are probabilistic finite state machines in which states correspond to POS tags, and observations to words (Zin and Thein 2009). HMMs are often used alongside the Viterbi algorithm (Viterbi 2006), a dynamic algorithm used to discover the hidden state path (Cahyani and Vindiyanto 2019), that is, the most probable POS tags for a given textual sequence. Support Vector Machine (SVM) is a machine learning algorithm for binary classification that, in its simplest form, learns a linear hyperplane that separates one set of examples (so called positive examples) from another set (negative examples) (Antony,

Mohan, and Soman 2010). Even though it was originaly designed for binary classification, it was adapted to work with multiple classes as well (Hsu and Lin 2002). In SVM based POS tagging, multiclass classification is tackled by taking one POS tag at a time as a positive class, and the rest as negative (Fernando et al. 2016). Conditional Random Fields (CRFs) (Lafferty, McCallum, and Pereira 2001) are used to build probabilistic models that are able to segment and label the sequence data (Pallavi and Pillai 2014), and are in fact random fields globally conditioned on the observations that might range over natural sentences (Lafferty, McCallum, and Pereira 2001). They are often used in various NLP tasks (e.g., for NER identification in Patil, Patil, and Pawar [2020]) and not just for POS tagging. Stochastic approaches to POS tagging also commonly make use of *n*-gram analysis (e.g., Mittal, Sethi, and Sharma 2014), which can be defined as the deconstruction of text into smaller parts containing *n* words (e.g., the sentence "over the mountains" could be split into "over the" and "the mountains" in 2-gram or bigram analysis). This kind of analysis enables the examination of context because it allows for words to be analyzed alongside their neighbors, which in turn provides insight into the kind of words that usually appear together.

Hybrid approaches to POS tagging encompass methods that combine rule-based and stochastic approaches to POS tagging.

6.1.5 NER Identification. NER (Named Entity Recognition) identification can be defined as the process of determination of words that denote the names of people, places, organizations, dates, common abbreviations, and so forth. It represents one of the crucial steps in knowledge extraction and in the construction of semantic networks and knowledge graphs commonly used in artificial intelligence. Lately, NER identification methods are commonly deep learning based (e.g., Lample et al. 2016; Yan, Jiang, and Dang 2021; Cui et al. 2021), but also encompass methods based on hand-crafted rules (e.g., Riaz 2010; Studiawan, Hasan, and Pratomo 2023), CRFs (e.g., Chen et al. 2019; Patil, Patil, and Pawar 2020), HMMs (e.g., Azarine, Arif Bijaksana, and Asror 2019), and various combinations of different methods (e.g., Jin et al. 2019; Drovo et al. 2019; Yi et al. 2020).

It is important to note that many NER identification methods for widely spoken languages exist, but NER identification is still a challenging problem for low-resource languages. Ancient unknown languages unfortunately fall into this latter category, which subsequently makes the associated scripts extremely difficult to decipher. Automatic analysis of low-resource languages is an ongoing problem that has gained momentum in recent years (Haddow et al. 2022; Costa-jussà et al. 2022). It is our belief that the advancements in this field will have a direct correlation with the advancements in computational decipherments of ancient scripts. This is somewhat supported by the fact that even Michael Ventris used a process similar to NER identification in the decipherment of Linear B when he assumed that certain Linear B words indicated the names of places on Crete (Mycenaean Epigraphy Group 2023). Granted, his decipherment process did not involve computational models that are the main focus of our investigation, but parallels can still be drawn between his research and the current state-of-the-art models aimed at the decipherment of ancient scripts.

6.2 Processing of Textual Information

Processing of textual information is usually based on traditional machine learning methods, deep learning models, or a combination of both.

Traditional machine learning methods include well-known approaches such as HMMs, SVMs, and CRFs. These models are statistical learning methods that were previously discussed in Section 6.1 in terms of their use in pre-processing tasks of NLP.

Artificial neural networks are digital models capable of, more or less successfully, mimicking the internal workings and decision making procedures of a human brain. They are commonly associated with the term **deep learning**, which can be defined as a methodological toolkit for building multilayer neural networks (Saxe, Nelli, and Summerfield 2021). Deep learning based models have gained quite a momentum in the decipherment of ancient scripts in recent years. On one hand, they are able to model highly complex functions (Goodfellow, Bengio, and Courville 2016), but on the other hand they require large amounts of training data and computational resources. The amount of training data required for deep learning based models poses a significant challenge in the decipherment of ancient scripts, since that data usually does not exist in the required quantities nor is it digitalized. When it comes to ancient undeciphered scripts, the types of digital training data associated with them are either visual (i.e., digital photographs or renderings of tablets containing ancient inscriptions), textual (i.e., digitalized inscriptions), or auditory (i.e., sounds associated with symbols used in a particular script—but this type of data is extremely rare and usually converted to the form of phonetic transcriptions). The available types of training data for a particular undeciphered script have a significant impact on the selection of an artificial neural network type and model best suited for the automatic undecipherment task. Convolutional neural networks (CNNs) are preferred for visual data processing, while textual data is usually processed by artificial neural networks built with Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), Gated Recurrent Units (Cho et al. 2014), and so on. An important breakthrough in deep learning was the introduction of the attention mechanism by Bahdanau, Cho, and Bengio (2014). This attention mechanism generally improved the performance of the basic encoder-decoder architectures in artificial neural network translation systems that previously decreased with the increase in length of the input sentence. This breakthrough led further to Transformer models in 2017 (Vaswani et al. 2017) that are used today as a main architecture for large language models. Large language models such as Generative Pre-trained Transformers (GPTs) (Radford et al. 2018) (on which ChatGPT [OpenAI 2023] is based), BART (Lewis et al. 2020), PaLM 2 (Anil et al. 2023), and Phi-2 (Javaheripi and Bubeck 2023) represent a major breakthrough in NLP and a giant step towards general artificial intelligence.

Despite the significant progress made in NLP in recent years, automatic decipherment of ancient scripts still remains a challenging task. There are, however, some positive changes occurring as well. Open source models, solutions, and data sources are becoming increasingly common and therefore facilitate further research in NLP and automatic decipherment attempts. For example, the source code for the Ithaca deep neural network for the textual restoration and geographical and chronological attribution of ancient Greek inscriptions (Assael et al. 2022a) is available at Assael et al. (2022b). Additionally, an increase in digital data that can be used for training NLP models oriented towards ancient inscriptions has been noted recently. For example, the Perseus Digital Library containing Greek and Roman text collections (Crane 2023a,b) and searchable libraries of Linear A inscriptions presented in Salgarella and Castellana (2021a,b) and Hogan (2023) are available online.

Since the automatic decipherment of ancient scripts is a relatively novel area of research, and since not many methods that tackle this problem exist, it cannot be stated with certainty whether it is better to use traditional or deep learning based approaches. Currently, a combination of both methods is recommended.

If the reader is interested in a more thorough and in-depth review of statistical machine learning and deep learning methods, we would suggest the following: a detailed overview of traditional machine learning methods can be found in Bontempi (2021), while a survey presented in Sommerschield et al. (2023) gives a detailed overview of machine learning methods and approaches for tackling different aspects of ancient language research.

6.3 Evaluation

Traditional metrics used for measuring the performance of NLP models are BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2002) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin 2004), but these metrics have been shown to suffer from low correlation with human judgment (Blagec et al. 2022). Many submetrics related to BLEU and ROUGE have been proposed over the years (Graham 2015; Blagec et al. 2022), as well as many completely novel metrics (e.g., BERTScore; Zhang et al. 2020). However, it is important to note that it is currently impossible to construct or design a completely automatic method that requires no human intervention for evaluating every natural language processing model in existence, and this is especially true for large language models associated with generative artificial intelligence (e.g., GPT-3; Brown et al. 2020). When it comes to ancient undeciphered scripts, it is just as difficult to evaluate and score the results of their computational analysis since in many cases there is nothing to compare them against, and their evaluation remains an open problem. In the case of Bronze Age Aegean and Cypriot undeciphered scripts, the obtained results might be compared to other geographically close known scripts from a similar time period in order to try and evaluate them, and for this very purpose we discuss these scripts in Section 8. If the reader is further interested in the evaluation methods of common natural language processing tasks (that each come with their own set of challenges), we encourage them to consult Blagec et al. (2022) and Lee et al. (2023) for additional details.

7. Computational Approaches to Deciphering Ancient Scripts

Computational approaches to deciphering unknown scripts, as opposed to the standard epigraphical and paleographical decipherments, started appearing in the late 1990s, with works such as Knight and Yamada (1999).

Current computational approaches to deciphering unknown scripts can be divided into two categories: (1) approaches aimed at partial or complete translation of unknown text to one or more of the known languages, and (2) approaches aimed at the discovery of novel information about the undeciphered script that might aid in the future partial or complete translation. Most of the existing computational approaches to deciphering unknown scripts fall into the second category, and these approaches include things like cognate or related word detection, closest known language detection, automatic dataset augmentations, statistical analysis of ancient texts, phonetic decipherments and transcriptions, word alignments, and so forth.

Since there are not many existing computational approaches to deciphering ancient scripts (not just regarding the Bronze Age Aegean and Cypriot scripts, but in general), we decided to divide them into Aegean and Cypriot approaches (discussed in Subsection 7.1), and non-Aegean and non-Cypriot approaches (discussed in Subsection 7.2). Finally, we also discuss computational simulations of traditional decipherment methods in Subsection 7.3.

7.1 Existing Approaches for the Decipherment of Aegean and Cypriot Scripts

In this section we present an overview of computational approaches regarding the decipherment of Aegean and Cypriot scripts. Existing approaches for the computational decipherment of these scripts can be loosely divided into those mostly based on the analysis of individual symbols, those mostly based on the analysis of words or word parts, and those oriented towards the prediction of missing or damaged symbols and dataset augmentation. These categories do not have hard boundaries as many of the existing approaches for the decipherment of Aegean and Cypriot scripts combine elements from multiple categories, and their assignment to a particular category is therefore rather fuzzy.

Figure 5 shows a chronological overview of computational approaches for the decipherment of Bronze Age Aegean and Cypriot scripts alongside their brief descriptions and indications regarding the script or scripts they are focused on. The figure does not include information regarding the Archanes script and the Archanes formula for the simple reason that we did not manage to find any papers focused on their computational decipherment. From Figure 5 it is clear to see that the vast majority of the computational decipherments of Bronze Age Aegean and Cypriot scripts revolves around Linear A and Linear B syllabaries, and that the other scripts are not very well represented in this emerging research field. Furthermore, there is an obvious lack of a standardized digital dataset of Bronze Age Aegean and Cypriot inscriptions, as most of the existing approaches use internal datasets that are not available online.

7.1.1 Approaches Based on the Analysis of Individual Symbols. This category of approaches for the decipherment of Aegean and Cypriot scripts is oriented towards the comparison of individual symbols, and words, or word parts from one or more undeciphered scripts to the symbols, words or word parts from one or more deciphered scripts. These deciphered scripts are usually selected for this comparison on the basis of their geographical location (i.e., they were or are used in or near the general area where the undeciphered scripts were used), and on the basis of temporal proximity (i.e., they were used around the same time period as undeciphered scripts). This category of approaches encompasses several methods (Revesz 2015a, 2016a,b, 2017b,c,d; Daggumati and Revesz 2019, 2023; Corazza et al. 2021; Srivatsan et al. 2021; Corazza et al. 2022).

Revesz (2015a) presented a bioinformatics inspired analysis of the signs belonging to Linear A, Linear B, Phoenician, South Arabic, Greek, Old Hungarian, and Cretan hieroglyphic scripts. They compared the symbols from these scripts to each other in terms of visual and phonetic features and presumable meanings, and grouped them into four categories they named A, C, G, and T after the four DNA nucleotides (adenine, cytosine, guanine, and thymine). They used these categories to encode the seven scripts they analyzed in their research, and used ClustalW2 phylogenetic algorithms to construct their hypothetical evolutionary tree. ClustalW2 is a DNA or protein multiple sequence alignment program for three or more sequences (EMBL's European Bioinformatics Institute 2023). The authors in Revesz (2015a) reported that the results they obtained indicate that the seven compared scripts had a common ancestor. Their evolutionary tree consisted of three branches, the first one containing Linear A and Linear B scripts, the second one containing the Cretan hieroglyphic script, and the third one containing Old Hungarian, South Arabic, Phoenician, and Greek scripts.

Revesz (2016a) presented a bioinformatics inspired approach to determining the evolutionary language tree of the Cretan hieroglyphic, Linear A, Linear B, and Cypriot syllabaries, and the Greek, Phoenician, Old Hungarian, South Arabic, and

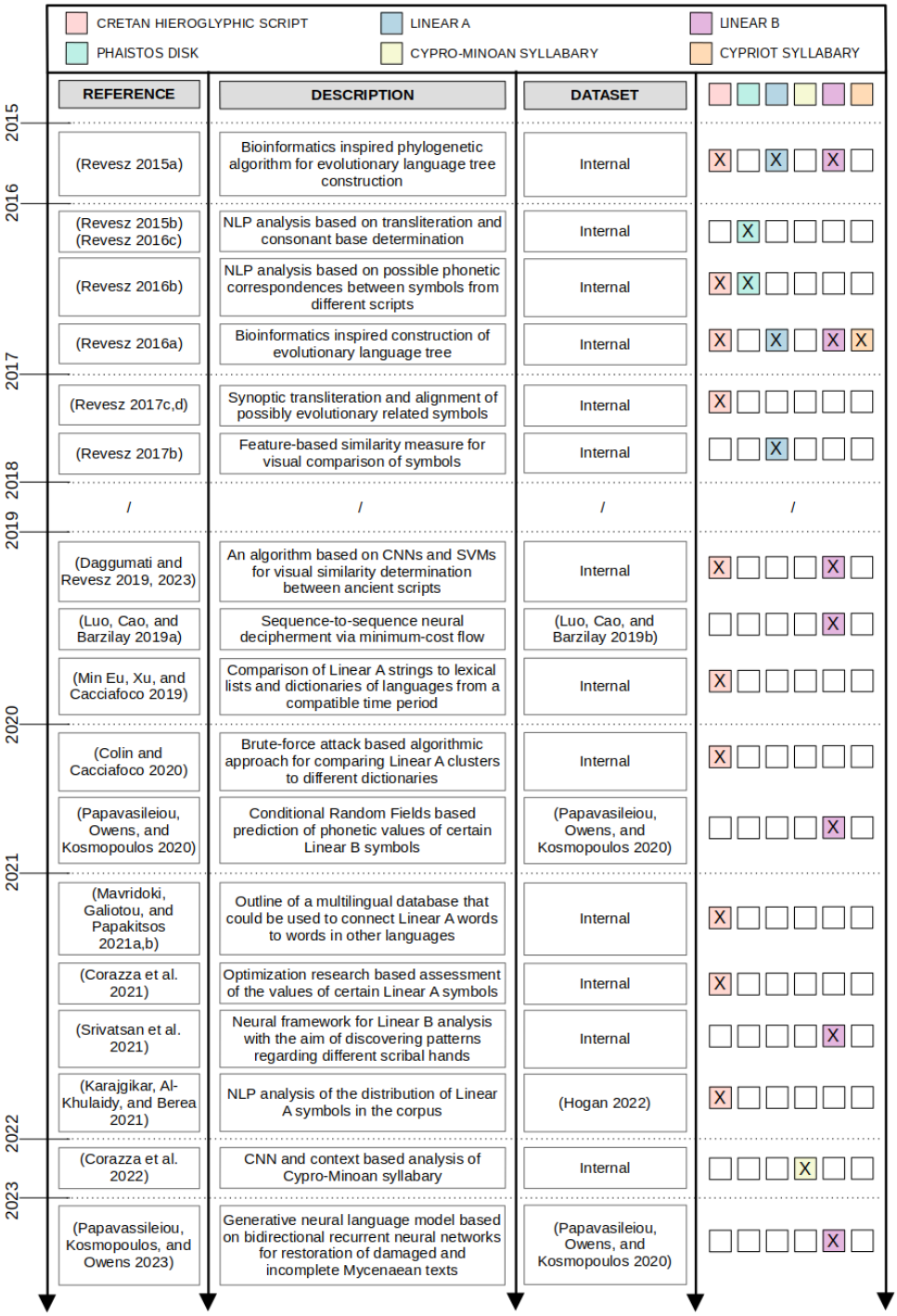


Figure 5
A chronological overview of computational approaches for the decipherment of Bronze Age Aegean and Cypriot scripts.

Tifinagh alphabets. The authors presented a table where they showed symbols from these scripts that have known phonemic values and compared them with each other. They proposed a measure for calculating the similarity of pairs of symbols in these scripts, and concluded that the Cretan hieroglyphic, Cypriot, Linear A, and Linear B syllabaries, and the Old Hungarian and Tifinagh alphabets belong to one language branch, while Greek, Phoenician, and South Arabic alphabets belong to a different language branch.

Revesz (2016b) proposed a computer-aided translation of the Cretan hieroglyphic script based on the possible phonetic correspondences between the symbols of the Cretan hieroglyphic script and the symbols of the Phaistos Disk. In their research they utilized findings from their previous work presented in Revesz (2016c), where they proposed a translation of the Phaistos Disk.

Revesz (2017c) presented a semi-automatic method for the translation of the Arkalochori Axe and the Malia Altar Stone inscriptions. In their method the authors used synoptic transliteration that aligned symbols from several scripts that are possibly evolutionarily related and whose phonetic values are at least partially known. This made it easier to guess potential phonetic values of symbols in the unknown script. The possibly related scripts that the authors used were Carian and Old Hungarian alphabets. The authors further extend their work on Malia Altar Stone analysis in (Revesz 2017d).

Revesz (2017b) proposed a feature-based similarity measure to visually compare symbols from different scripts. The proposed similarity measure compares symbols based on 13 different features—for example, whether the symbol contains some curved line or not, whether it contains parallel lines or not, and so on. They used this similarity measure to develop a new phonetic grid for the Linear A syllabary, and then construct an English-Minoan-Uralic dictionary. The algorithm that the authors developed can, for example, take a Linear A symbol as an input, and return its syllabic value.

Daggumati and Revesz (2019) and Daggumati and Revesz (2023) presented a method for discovering relationships (in terms of visual similarity) between ancient scripts. They used CNNs and SVMs to look for correlation between symbols in different scripts. The ancient scripts that the authors focused on included the Brahmi script (34 symbols were used from this script), Cretan hieroglyphic script (22 symbols were used), Greek alphabet (27 symbols were used), Indus Valley script (23 symbols were used), Linear B syllabary (20 symbols were used), Phoenician alphabet (22 symbols were used), Proto-Elamite script (17 symbols were used), and Sumerian pictographs (34 symbols were used). The dataset that the authors used included 900 images (780 for training and 120 for validation) for each symbol, and this large number of images was reached via image augmentation of hand-drawn symbols. The results that the authors obtained showed that the Linear B script is highly correlated with the Cretan hieroglyphic script.

Corazza et al. (2021) proposed a method based on optimization research for assessing the values of some of the more problematic numerical fraction signs used in the Linear A syllabary. Their method combined palaeographical, statistical, typological, and constraint-based computational approaches. In their work the authors listed all of the fraction signs presumably used in Linear A, and discussed different theories on their possible meanings. Even though Linear A and Linear B are inherently linked, the authors emphasized that the quantities of commodities are recorded differently in these two scripts (namely, Linear B did not use fractions), thus making the comparison between the two challenging. This comparison was possible, however, between Linear A fraction signs and the signs used in the "Egyptian systems (hieroglyphic, hieratic, 'Eye of Horus,' and demotic), Mesopotamian cuneiform (Old Akkadian/Old Babylonian

phases), Greek alphabetical, Coptic alphabetical, North African Fez/Rumi, the 'vulgar' Arabic system, and Indian Grantha." The research presented by the authors helped narrow down the range of possible values of Linear A fraction signs, and assisted in identifying some of the languages and scripts that may be useful in deciphering it.

Srivatsan et al. (2021) proposed a neural framework for Linear B script analysis and demonstrated it on the scribal hand representation learning from raw images of Linear B glyphs. The aim of their study was to develop an approach that could potentially be used in uncovering patterns associated with ways in which various scribes may have written the same symbol, and perhaps to even discover how the Linear B writing system evolved. The authors represented each Linear B glyph with two vector embeddings—one associated with the potential scribal hand that made it, and one representing the syllable the glyph is associated with. The vector embeddings associated with scribal hands are learned during the training phase of the convolutional neural network the authors used. In order to avoid overfitting due to the limited size of the dataset, the authors used image augmentation to increase the size of the dataset from 4,134 to 111,618 images. The augmentation types that they used consisted of dilation, erosion, and horizontal and vertical translation. The authors identified a total of 74 scribal hands associated with the images the model was trained on, and evaluated their results on baseline human generated scribal hand representations from Skelton (2008). The authors report that the results obtained by their method were closer to the baseline manual representations than the results obtained by an autoencoder model they compared their method against.

Corazza et al. (2022) proposed a method based on CNNs for classification of symbols belonging to the Cypro-Minoan syllabary. Their model was named $Sign2Vec_d$, and it represented a modification of the DeepCluster-v2 model (Caron et al. 2018, 2020). The authors used context in their approach and looked at what symbols usually appeared together in inscriptions. The model was trained on 2,899 Cypro-Minoan sign images obtained from 213 inscriptions. The results they obtained indicate that the Cypro-Minoan syllabary is a unitary script rather than three separate varieties, and that the differences between Cypro-Minoan symbols are mainly due to the material of the artefact on which they are located.

7.1.2 Approaches Based on the Analysis of Words or Word Parts. This category of approaches for the decipherment of Aegean and Cypriot scripts is oriented towards the comparison of words or word parts between undeciphered and deciphered scripts, in order to potentially discover some connections between them that could aid in the decipherment process. This category of approaches encompasses several methods (Revesz 2015b, 2016c; Luo, Cao, and Barzilay 2019a; Min Eu, Xu, and Cacciafoco 2019; Colin and Cacciafoco 2020; Mavridaki, Galiotou, and Papakitsos 2021a,b).

Revesz (2015b) and Revesz (2016c) presented a semi-automatic method for the translation of the Phaistos Disk by using its connections to other ancient languages and scripts, including the Proto-Finno-Ugric language and the Old Hungarian alphabet. Their method includes five steps: transliteration of the Phaistos Disk symbols, construction of the Proto-Finno-Ugric and Proto-Hungarian dictionary, determination of a consonant base for each word in the dictionary, determination of matches between transliterated text and dictionary words, and the formation and translation of sentences. Their work was based on Revesz (2015a) and the assumption that the sound changes within a word (during a longer time period) are similar to genetic mutations. According to the authors, the Phaistos Disk possibly represents an ancient sun hymn that might be connected to a winter solstice ceremony.

Luo, Cao, and Barzilay (2019a) proposed a method for the automatic decipherment of lost languages that is based on LSTM based sequence-to-sequence neural decipherment via minimum-cost flow. The goal of their model was to determine character-level correspondences between cognates or related words. They tested their method on the Linear B syllabary, where their method correctly translated 67.3% of cognates on a challenging Linear B dataset, and 84.7% on a noiseless, less challenging Linear B corpus. In the challenging Linear B dataset, roughly 50% of the Linear B were missing a cognate word in Greek. The dataset that the authors used in the development of their method is available at Luo, Cao, and Barzilay (2019b), and presents a modification of the one presented in Tselentis (2011).

Min Eu, Xu, and Cacciafoco (2019) presented an initial design of a computational method for comparing Linear A strings to lexical lists and dictionaries of languages from a compatible time period. Because their proposed design is still under development, not much in-depth information is available on it.

Colin and Cacciafoco (2020) proposed a "brute-force attack" based algorithmic approach for comparing Linear A clusters to different dictionaries of languages used in the relative chronological and geographical vicinity to Linear A. Their aim was to perform an exhaustive search for a language or a language family that the Linear A script might be related to. This search included the comparison between words from the dictionaries and lexical lists of different languages, and the list containing information on Linear A. Preliminary results that the authors obtained indicate that the links could exist between Linear A symbols and symbols found in languages such as Luwian, Thracian, Hamito-Semitic, and so forth.

Mavridaki, Galiotou, and Papakitsos (2021a) and Mavridaki, Galiotou, and Papakitsos (2021b) presented an outline of a software tool that could be used for understanding and learning Linear A. Their tool contains a multilingual database and could be used to connect Linear A words to related words in other languages. Since the software tool the authors presented is still under development, not much information on it has been released to the public.

7.1.3 Approaches Oriented Towards the Prediction of Missing or Damaged Symbols. This category of approaches for the decipherment of Aegean and Cypriot scripts is mostly oriented towards the prediction of missing or damaged symbols with the aim of text restoration and dataset augmentation. Since not many ancient texts associated with certain undeciphered languages exist, this could prove incredibly important for future researchers. This category of approaches encompasses several methods (Assael, Sommerschield, and Prag 2019; Papavassileiou, Owens, and Kosmopoulos 2020; Karajgikar, Al-Khulaidy, and Berea 2021; Papavassileiou, Kosmopoulos, and Owens 2023).

Assael, Sommerschield, and Prag (2019) proposed a model for ancient text restoration that recovers missing characters from damaged texts by the use of bidirectional LSTM based deep neural network. They named their model Pythia and trained it on Ancient Greek inscriptions dating from the seventh century BCE to fifth century CE. For each text restoration task Pythia returns 20 most probable outcomes (decoded by beam search) instead of just one. The Pythia model was expanded in Assael et al. (2022c), where the authors presented a deep neural network for textual restoration, geographical attribution, and chronological attribution of ancient Greek inscriptions and named it Ithaca. Ithaca achieves 62% accuracy on tasks revolving text restoration, but also helps historians in improving their accuracy from 25% to 72%. Additionally,

Ithaca can determine the geographical origin of ancient texts with 71% accuracy, and ascertain the timeframe in which they were created with an error range of ±30 years.

Papavassileiou, Owens, and Kosmopoulos (2020) presented a dataset of Mycenaean Linear B sequences. On one part of this dataset they conducted an experiment that aimed to predict the phonetic values of unidentified or illegible Linear B signs by using a linear-chain CRF based approach. The part of the dataset that was used in this experiment included tablets from Knossos that referred to lists of personnel. The authors plan on making the dataset public once it is completed and reviewed by experts in the field.

Karajgikar, Al-Khulaidy, and Berea (2021) proposed a method for exploring the Linear A dataset with the purpose of obtaining information on the distribution of symbols in the corpus. Their method consisted of many different natural language processing techniques, such as generating bigrams and trigrams, using weighted frequency for symbol prediction, word embedding, and so on. The framework of their proposed model included exploratory *n*-gram analysis and exploratory knowledge mining. Exploratory *n*-gram analysis revealed high levels of entropy within the probable words in Linear A dataset, higher than the levels commonly found in other languages. This indicated to the authors that Linear A may not represent a language at all. However, when the authors performed the analysis on the entire Linear A dataset and not just on probable words from the dataset, the entropy decreased. Results of the exploratory *n*-gram analysis were therefore inconclusive. Exploratory knowledge mining, on the other hand, was used to determine whether the symbols in the dataset represented words or ideograms. The authors compared *k*-Nearest Neighbors and naïve Bayes classifiers for that task, and ultimately chose the former as the final model.

Papavassileiou, Kosmopoulos, and Owens (2023) presented a generative neural language model based on bidirectional recurrent neural networks for restoration of damaged and incomplete Mycenaean texts (e.g., the ones written in Linear B). Their method could be used for augmenting already scarce inscriptions contained in Linear B datasets by means of predicting the missing syllables. The dataset of Linear B sequences that the authors used in their model is the one presented in their previous work in Papavassileiou, Owens, and Kosmopoulos (2020), but unfortunately unavailable online. In order to help improve the accuracy of their augmentation method, the authors manually extracted 4 rules from a subset of the Linear B dataset. This subset is called *series D*, and it consists of 1,100 tablets that contain information about sheep herds of Crete. The rules that the authors extracted from that subset are the results of the linguistic analysis and include information about common structure of Linear B tablets from series D (e.g., the common order of toponyms and personal names on texts inscribed on tablets). Finally, the authors released a table containing inscriptions from damaged tablets alongside their most probable augmentation generated by the proposed model.

7.2 Existing Approaches for the Decipherment of Non-Aegean and Non-Cypriot Scripts

In the interest of covering a greater selection of algorithms that might be used for the decipherment of ancient scripts, in this section we present a chronological overview of computational approaches regarding the decipherment of non-Aegean and non-Cypriot scripts. As opposed to the Aegean and Cypriot scripts, these scripts may use non-syllabic writing systems and may come with their own unique sets of challenges that fall outside of the scope of this article.

There are several existing approaches for the computational decipherment of non-Aegean and non-Cypriot scripts (Knight and Yamada 1999; Knight et al. 2006; Terras 2006b; Snyder and Barzilay 2008; Snyder 2010; Snyder, Barzilay, and Knight 2010; Snyder and Barzilay 2010; Berg-Kirkpatrick and Klein 2011; Kim and Snyder 2013; Currey and Karakanta 2016; Deri and Knight 2016; Luo et al. 2021). We divide these approaches into three categories: (1) those based on the analysis of cognate words, (2) those based on the digitalization of the papyrology process, and (3) those based on using information gathered from high-resource languages. These categories intertwine and do not have hard set boundaries.

7.2.1 Approaches Based on the Analysis of Cognate Words. This category of approaches encompasses methods based on the discovery and analysis of related words from different languages and includes several approaches (Snyder, Barzilay, and Knight 2010; Berg-Kirkpatrick and Klein 2011; Luo et al. 2021).

Snyder, Barzilay, and Knight (2010) proposed a method for the automatic decipherment of lost languages that is based on a non-parametric Bayesian framework. Their method required a non-parallel corpus in a known related language to produce alphabetic mappings and identify cognate words, and was based on the assumption that the writing system of the language trying to be deciphered was alphabetic in nature. The authors concentrated on Ugaritic and Hebrew languages. When using Hebrew as a related known language to Ugaritic, their method accurately mapped 29 of 30 Ugaritic letters, and accurately translated 60.4% of all cognate words. Their proposed method was elaborated upon in additional detail (Snyder 2010; Snyder and Barzilay 2010).

Berg-Kirkpatrick and Klein (2011) proposed a method for the detection of cognate words between two scripts written in an alphabetic writing system. They formulated the problem they were trying to solve as a combinatorial optimization problem where coordinate descent procedure was used to discover solutions. Their method incorporated both a matching between alphabets and a matching between lexicons, and was evaluated on four datasets where it achieved promising results. The languages that the method was evaluated on included English, Italian, Spanish, Portuguese, Hebrew, and Ugaritic.

Luo et al. (2021) proposed a generative framework for deciphering undersegmented ancient scripts using a phonetic prior. The inputs to their model include an undersegmented inscription in an undeciphered language written in an alphabetic writing system, and a vocabulary in a deciphered language also written in an alphabetic writing system. The aim of their model is to extract spans from the undeciphered text and match them to cognates in a deciphered text. The authors used International Phonetic Alphabet (IPA) embeddings and represented each symbol of a known language with a vector of its phonological features. The authors also proposed a measure for language closeness that can be used to discover languages that are close or related to the undeciphered language. The authors showed that their method accurately detected Gothic and Proto-Germanic deciphered languages as related, Ugaritic and Hebrew deciphered languages as related, and Iberian undeciphered language as probably not related to any of the languages they compared it against. They also concluded that Iberian is more similar to Basque (also undeciphered) than to any of the other compared languages. Since the selection of compared languages that they used was rather small (they used Latin, Spanish, Turkish, Hungarian, Proto-Germanic, and Basque), these findings could perhaps be expanded upon in the future by the inclusion of a greater number of languages in their linguistic analysis.

7.2.2 Approaches Based on the Digitalization of the Papyrology Process. This category of approaches encompasses methods based on the digitalization of the traditional papyrology process, using the following approach.

Terras (2006b) proposed a computational model of the papyrology process that can aid in the complicated procedure of reading very damaged ancient documents. They demonstrated their method, based on image processing and image analysis via the adapted GRAVA (Grounded Reflective Adaptive Vision Architecture) architecture (Robertson 1999, 2001), on the Vindolanda tablets discovered in the ruins of ancient Roman fort Vindolanda located in Chesterholm in Northern England. Although not initially proposed for the decipherment of unknown texts, their model behaves in a similar manner to the expert papyrologists and can potentially accelerate either the decipherment process, or at least the construction of digital datasets of unknown texts and inscriptions. The method proposed in Terras (2006b) was further elaborated upon in Terras and Robertson (2005) and Terras (2006a).

7.2.3 Approaches Based on Using High-resource Languages as a Base. This category of approaches encompasses methods based on using information gathered from high-resource languages for the analysis of possibly related low-resource languages. For the purposes of model evaluation and testing, researchers sometimes use high-resource languages as if they were low-resource languages and then try to "decipher" them. This category of approaches has been used by many papers (Knight and Yamada 1999; Knight et al. 2006; Snyder and Barzilay 2008; Kim and Snyder 2013; Currey and Karakanta 2016; Deri and Knight 2016).

Knight and Yamada (1999) proposed a method for the phonetic decipherment (or text-to-speech conversion) of unfamiliar scripts written in a known language (as in the case of Aegean Linear B). They used a bidirectional sound-to-character sequence finite-state transducer, and used Expectation-Maximization algorithm to determine sound-character mapping probabilities. In their work the authors concentrated on Spanish, Japanese, and Chinese languages. Even though they did not test their algorithm on any ancient unknown script, they did conclude that the decipherment is possible even with only limited knowledge of the language behind that script. This is promising because oftentimes not much information is known about a language behind an unknown script, and this is certainly true in the case of undeciphered Bronze Age Aegean and Cypriot scripts.

Knight et al. (2006) proposed a method for automatic decipherment of ciphertexts into plaintexts that is based on the Expectation-Maximization algorithm. The authors particularly focused on phonetic decipherments, and divided them into two separate categories: regular phonetic decipherments (when the language behind the ciphertext is known), and universal phonetic decipherments (when the language behind the ciphertext is unknown). Ciphertexts can in the latter case be regarded as texts that the researchers have trouble deciphering, such as Linear A inscriptions and most of the texts and documents associated with the Minoan civilization. The authors built phoneme n-gram databases for 80 different languages, and showed that when compared to these databases, the encoded input text written in the Spanish language and encoded with a simple substitution cipher can successfully be recognized as being written in Spanish (and if not in Spanish, then the model predicts Galician, Portuguese, or Kurdish languages as being the closest matches to the encoded text). The work presented in Knight et al. (2006) can therefore be extremely helpful in discovering more information about ancient scripts written in unknown languages.

Snyder and Barzilay (2008) presented a non-parametric, unsupervised, and hierarchical Bayesian model that induces morpheme segmentations of multiple selected languages and concurrently identifies cross-lingual morpheme patterns. The authors concentrated on Arabic, Hebrew, Aramaic, and English languages. The motivation behind their work was to discover connections between different languages and determine whether those connections could prove useful in automatic language analysis.

Kim and Snyder (2013) proposed a method for the prediction of consonants and vowels for an unknown language and alphabet. They performed posterior inference over 503 languages in order to obtain information that would allow them to discover general linguistic patterns. In order to obtain that information, they used HMMs and made 3 assumptions: (1) each language has an unobserved set of parameters that can explain its observed vocabulary, (2) each language-specific set of parameters has an unobserved common prior shared between a cluster of related languages, and (3) each of those clusters derives its parameters from a common prior of all language groups. The motivation behind their research was to discover, for a given unknown language and alphabet, a group of languages they are related to. This would provide them with certain assumptions on common linguistic patterns which they could use in the prediction of consonants and vowels. Even though their research was limited to alphabets, they plan on expanding it to include other writing systems as well.

Currey and Karakanta (2016) proposed a method for lessening the problem of a lack of training data for low-resource languages[3] by augmenting them with data from related high-resource languages. In their work, the authors treated Spanish language as if it were a low-resource language, and used Italian and Portuguese as related high-resourced languages. Although their method did not show that language modeling for a low-resource language can be improved by using information from related high-resource languages, it did show promising results in the field of statistical machine translation.

Deri and Knight (2016) proposed a grapheme-to-phoneme model trained on high-resource language(s) and applied to related low-resource language(s). Grapheme-to-phoneme models convert written words into pronunciations that are usually represented in IPA. The dataset on which the authors trained the model consisted of more than 650,000 word-pronunciation pairs from more than 500 languages. The authors made use of an online repository of cross-lingual phonological data named Phoible (Moran, McCloy, and Wright 2014).[4] Phoible is composed of language phoneme inventories that contain sets of phonemes represented in IPA, and each phoneme also contains a unique feature vector that represents its phonological features.

7.3 Computational Simulations of Traditional Decipherment Methods

Traditionally, decipherments of ancient scripts are based on epigraphy, paleography, and linguistics. Michael Ventris and his colleagues who jointly contributed to the decipherment of Linear B did not have computer resources to help them in their

3 Low-resource languages are generally considered as languages for which not much data exists that can be used in natural language processing applications; high-resource languages are their exact opposites.

4 The version of the Phoible repository that the authors in Deri and Knight (2016) used is now perhaps out-of-date, as an expanded version of the repository, known as Phoible 2, is available at Moran and McCloy (2019).

decipherment process, but were still able to successfully achieve their goal. In this section of the article we summarize the traditional Linear B decipherment process and propose automatic methods that could be used to simulate it.

According to the Mycenaean Epigraphy Group (2023), the Linear B decipherment process included the following:

1. determination of Linear B's writing system—it was a syllabary based on the number of phonetic symbols, but contained a smaller fraction of logograms as well;
2. determination of an overall content of Linear B inscriptions—in certain cases it was possible to infer that the inscriptions represented administrative documents based on easily recognizable logograms that provided the much needed contextual information;
3. establishment of the definitive list of Linear B symbols and their variants with the same phonetic values—this analysis was performed by Emmett L. Bennett Jr.;
4. statistical analysis of Linear B words that only slightly differed from one another—this analysis was performed by Alice Kober and showed that the language behind Linear B was an inflected language—that the words could change depending on, for example, gender;
5. educated guesses—Michael Ventris made a series of educated guesses about Linear B script, such as hypothesizing that certain Linear B words could represent the names of real places on Crete, that vowels are usually found in the beginning of the words in syllabic scripts, and that consonants followed by vowels usually appeared in the middle. Ventris then created a table depicting the symbols and their possible phonetic values, and through trial and error realized that the language behind Linear B was a form of archaic Greek.

To simulate the above-mentioned steps one could use a number of different computational methods discussed in Section 6. Here we present a general outline of a computational approach to five traditional decipherment steps outlined above. In all five steps we assume that there is enough data on which the algorithm can be trained or analysis performed, but sadly this is not always the case.

7.3.1 Determination of a Writing System Associated with an Unknown Script. Knowledge about the writing system used in an unknown script is extremely important. A decipherment process associated with an alphabet based writing system would be very much different from one based on a logosyllabary. Sometimes, a good guess about the type of writing system used in a certain unknown script can be based upon similar or possibly related scripts (this may not be known), on a number of unique symbols used in an unknown script (this may also be unknown), or on a statistical language analysis of available texts and an examination of their internal structure.

Usually, alphabets have a small number of unique symbols while logosyllabaries have many more. A solid line that could be drawn between different types of writing

systems unfortunately does not exist, and many scripts exhibit the characteristics of multiple types of writing systems instead of just one. Computational approaches associated with these processes could include expert-based rules regarding the number of unique symbols found in unknown scripts, a statistical comparison with scripts whose writing systems are already known, and a computational expert-based statistical analysis of grammatical structure.

7.3.2 Context Determination. In order to determine the overall context of an undeciphered text, one must first look at the location at which the text was found. If the text was found in the ruins of a religious monument, one might conclude that the text itself had a religious meaning. If the text was found on or near the remnants of an old market, perhaps it contained numbers and fractions associated with the prices of goods once offered at that market. However, since the original location of the text cannot always be determined, one might use image processing based search and recognition of certain pictograms and logograms in order to gain insight of the overall context of the text. If, for example, an automatic recognition system recognizes a pictogram or a logogram resembling a person, it might conclude that the symbols near it might represent that person's name. This visual recognition system might use traditional image processing methods (e.g., corner detection, edge detection, and various other filters) or convolutional neural networks trained on logograms with already known meanings. The same algorithm could also be modified to search for similarities between pictograms and logograms in different scripts, and try to determine the relationships (if any) between them.

7.3.3 Vocabulary Construction. Vocabulary construction is the most important step of the decipherment process since many of the other steps depend on it, but it can also be the most challenging one to realize. Symbols can have many different variations and in order to uncover them all, a statistical analysis of the text must be performed. Once obtained, these results can then be compared to the grammatical features of other languages in order to uncover the relationship (if any) between them.

7.3.4 Statistical Text Analysis. Statistical analysis of words in an undeciphered text assumes that the text can be segmented into words. This is not always the case, however, as the punctuation marks might not even exist or they might be unknown. If the words are able to be extracted from the text, a simple calculation of an edit distance between the extracted words will locate similar words and provide a measure of their similarity. Further grammatical analysis of similar words can try to discover whether the language behind the script is inflectional or not.

7.3.5 Educated Guesses. This step of the decipherment process of ancient scripts is possibly the most difficult one to simulate via a computational method. It depends on the knowledge from many different areas (historical, geographical, grammatical, linguistic, etc.), and on a neural process in which that knowledge is ultimately combined. In order to even begin to realize this step of the decipherment process, a construction of a generalized database that incorporates all of that knowledge is required. How this should be performed, however, and what data should be included in it, is still an open question.

7.4 Final Remarks on the Computational Decipherment of Ancient Scripts

It can be concluded from previous sections that the computational research on ancient unknown scripts is a complex and challenging field. With the advancements in computer science, and especially in artificial intelligence and deep learning models, this field continues to evolve. Our final remarks on the reviewed approaches for the computational decipherments of ancient scripts are as follows:

- There is an obvious lack of digital data and corpora associated with the computational decipherment of ancient scripts. Datasets currently linked to ancient unknown scripts are not standardized among the researchers, and in return this makes various decipherment models and associated results difficult to compare.
- The datasets associated with ancient unknown scripts are generally small and require extensive computational augmentation. However, since the methods used for these augmentations are not standardized, even if different researchers start working with the same dataset, that dataset might change drastically after augmentation.
- Different methods for the decipherment of ancient scripts can start with different assumptions, and therefore arrive at completely different conclusions. This is especially true for cases when researchers assume that the ancient unknown scripts are related to certain languages and scripts, as these can widely differ.
- Computational approaches for the decipherment of ancient scripts usually involve comparison with chronologically and geographically related languages and scripts. This comparison is usually based on the analysis of visual and phonological features of symbols and words, with the aim of identifying those that might be related and have similar meanings.
- Deep learning based models for the computational decipherment of ancient scripts are becoming increasingly common, and are starting to outnumber the methods based on traditional machine learning (e.g., SVM or HMM). This is not surprising, however, as large language models and generative artificial intelligence are capable of making connections between different data points that the traditional machine learning algorithms are simply not designed for. This makes deep learning models better at finding hidden connections between different languages and scripts, which can in turn greatly impact the field of computational decipherment of ancient scripts.

Computational decipherment of ancient scripts is an incredibly interesting and rapidly growing field. With the advancements in computer science, we believe it will one day be possible to decipher not only Bronze Age Aegean and Cypriot scripts, but many other unknown scripts as well. We hope that this will lead to new insights about the lives of the people who used these scripts, as well as to additional information about our own history as well.

8. Related Scripts and Languages

One of the most important things in deciphering unknown ancient languages is to try to determine their closest relative(s), as this could have a positive impact on the resolution of challenges outlined in Section 4. The importance of using related languages to enhance statistical language models is not new, and it was already emphasized in previous work (Currey and Karakanta 2016; Pourdamghani and Knight 2017; Karakanta, Dehdari, and van Genabith 2018; Pourdamghani and Knight 2019; Mavridaki, Galiotou, and Papakitsos 2021b). Tóth (2007) also states that the distribution of many typological features of languages is not random but restricted to a relative geographical area. Since the language(s) behind the Bronze Age Aegean and Cypriot scripts are mostly unknown, it is helpful to look at the history of those scripts and the people that used them in order to gather more information about them and connect them to possible related scripts and languages.

British archaeologist Sir Arthur J. Evans named the Bronze Age culture of Crete Minoan (Chadwick 1987). Even though Evans hypothesized that the Minoans were refugees from Northern Egypt (Callaway 2013), the ancestry of the Minoan civilization is still a highly debated topic and no clear stand on this issue exists among researchers. Recent analysis of mitochondrial DNA taken from the Minoan osseous remains from a cave ossuary in the Lassithi plateau of Crete that was presented in Hughey et al. (2013) refutes the Evans hypothesis of Minoan North African origin, and supports the hypothesis of autochthonous development of the Minoan civilization by the descendants of the Neolithic settlers of Crete who probably arrived there around 9,000 years ago from Anatolia and/or the Middle East (Hughey et al. 2013). Research presented in Hofmanová et al. (2016) demonstrated that a direct genetic link exists between the Mediterranean and Central European early farmers and those of Greece and Anatolia. It should be emphasized that the authors in Hofmanová et al. (2016) concentrated on paleogenomic data obtained from northern Greece and northwestern Turkey, and not on Crete or Cyprus. Additionally, research presented in Omrak et al. (2016) shows a direct link between Anatolia and the early European Neolithic gene pool. Other theories on the Minoan origins exist as well. For example, in Lazaridis et al. (2017) the Minoan ancestry is linked to Anatolian and Aegean Neolithic farmers, and to the ancient populations related to those of Caucasus and Iran; in Revesz (2019) the Minoan ancestry is linked to Anatolia, Danube Basin, and the Black Sea littoral area; and in Revesz (2021) the Minoan ancestry is linked to Northern Greece, Anatolia, Caucasus, and the Balkans (mostly the Danube Basin). Previous attempts at linking the Bronze Age Aegean and Cypriot scripts to other possibly related scripts can also be found (Revesz 2017b; Schrijver 2019; Revesz 2020).

Languages used during the Bronze Age in relative proximity to Greece and Cyprus that we will focus on in this section include the Sumerian, Ancient Egyptian, Eblaite, Akkadian, Hattic, Hittite, Palaic, Luwian, and Phoenician languages. These languages were mostly used in certain places close to the Mediterranean Sea, especially Mesopotamia, Ancient Egypt, Anatolia, and Phoenicia. The timeline of these languages is summarized in Table 1, and discussed and referenced in detail in sections 8.1–8.4. Additionally, in section 8.5 we discuss geographically close languages that became prominent after the disappearance of Bronze Age Aegean and Cypriot scripts, as these languages could have potentially been influenced by the disappeared Aegean and Cypriot scripts, and could perhaps be helpful in their decipherment.

Figure 6 shows the approximate locations of Mesopotamia, Ancient Egypt, Anatolia, and Phoenicia. For the creation of the map we used QGIS geospatial software

Table 1
Bronze Age languages used in relative proximity to Crete and Cyprus

Language	Linguistic classification	Script	Geographical region	Approximate timeframe
Sumerian	Language isolate	Cuneiform	Mesopotamia	4000-2000 BCE
Old Egyptian	Afro-Asiatic → Ancient Egyptian	Egyptian hieroglyphs	Ancient Egypt	2700-2200 BCE
Eblaite	Afro-Asiatic → Semitic → East Semitic	Adapted cuneiform	Syria	2450-2350 BCE
Akkadian	Afro-Asiatic → Semitic → East Semitic	Adapted cuneiform	Mesopotamia	2350 BCE - first century CE
Middle Egyptian	Afro-Asiatic → Ancient Egyptian	Egyptian hieroglyphs	Ancient Egypt	2200-1800 BCE
Hattic	Language isolate, possibly Uralic → Finno-Ugric → Ugric	Not recorded by its native speakers; known from Hittite cuneiform texts	Anatolia	late third to mid-second millennium BCE
Kalasmic	Indo-European → Anatolian	Cuneiform	Anatolia	1650-1200 BCE
Hittite	Indo-European → Anatolian	Cuneiform	Anatolia, Syria	1650-1180 BCE
Palaic	Indo-European → Anatolian	Cuneiform	Anatolia	sixteenth to twelfth century BCE
New Egyptian	Afro-Asiatic → Ancient Egyptian	Egyptian hieroglyphs	Ancient Egypt	1580-700 BCE
Luwian	Indo-European → Anatolian	Adapted cuneiform, Anatolian hieroglyphs	Anatolia, Syria	1500-700 BCE
Phoenician	Afro-Asiatic → Proto-Semitic → West Semitic → Central Semitic → Northwest Semitic → Canaanite	Phoenician alphabet	Lebanon and many other Mediterranean coastal areas	twelfth century BCE - 196 CE

(QGIS Development Team 2023). A further review of computational natural language processing approaches to similar and related languages, language varieties, and dialects can be found in Zampieri, Nakov, and Scherrer (2020).

8.1 Mesopotamian Languages

The historical region of Mesopotamia approximately includes the area that is now eastern Syria, southeastern Turkey, and most of Iraq (Frye et al. 2023).

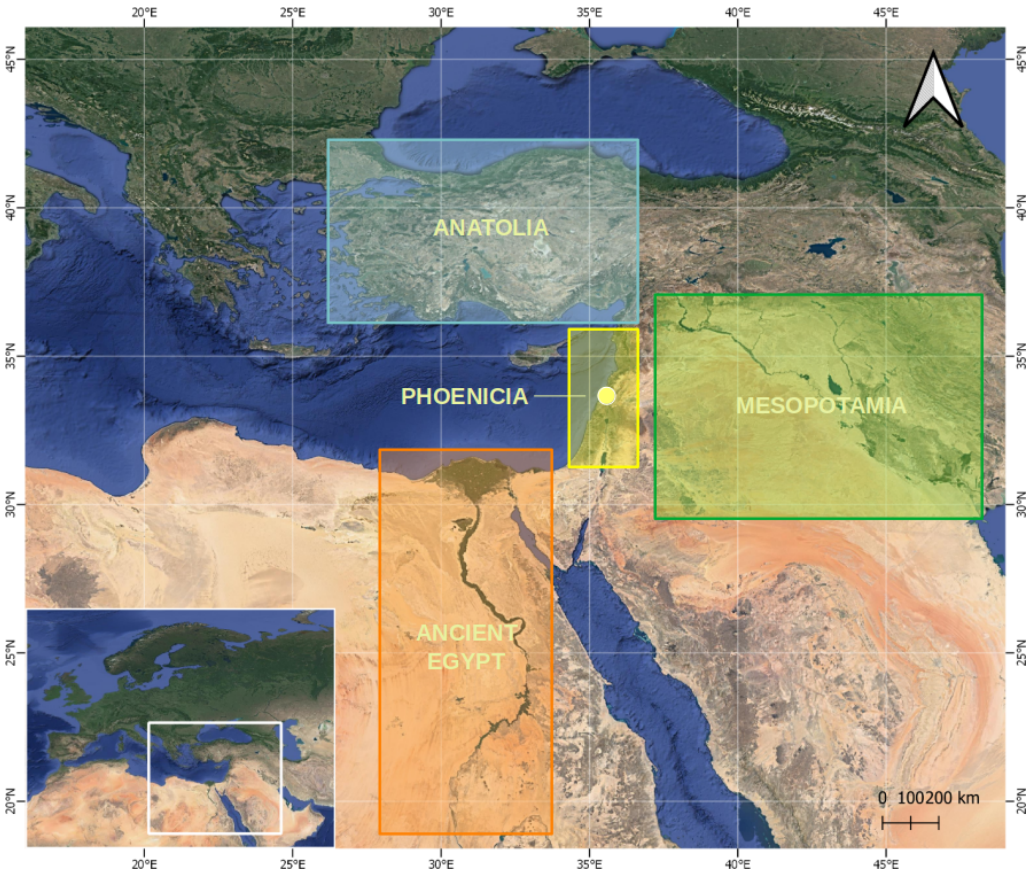


Figure 6
Approximate locations of Mesopotamia, Ancient Egypt, Anatolia, and Phoenicia.

Mesopotamian languages were written mostly in a cuneiform script. According to Kudrinski (2016), cuneiform script was invented in the late fourth millennium BCE for writing one of the languages of southern Mesopotamia, supposedly Sumerian. The Sumerian language (c. 4000-2000 BCE [Hajiyeva 2019]) is considered to be a language isolate (Pereltsvaig 2021), which means it does not appear to share an ancestral language with any other known language. It was written in the Sumerian cuneiform script (c. 3400 BCE - 75 CE [Gutherz et al. 2023]). The paleographic links between the Sumerian language and the Aegean scripts (mainly Linear A and Linear B) are outlined in Papakitsos and Kenanidis (2015).

Two additional Mesopotamian languages that might be related to the Aegean scripts are Eblaite and Akkadian. These languages were written in a script adapted from Sumerian cuneiform (Pereltsvaig 2021). Eblaite was spoken in northeast Levant or present-day Syria (Gordon 1997) (as cited in Kitchen et al. 2009), and is attested only from the Ebla archives (c. 2450-2350 BCE) (Pearce 2010). Akkadian language (c. 2350 BCE – first century CE [Seri 2010]) was spoken in Mesopotamia (Beckman 1983). Eblaite and Akkadian belong to the Afro-Asiatic → Semitic → East Semitic language family (Hetzron 2009).

8.2 Ancient Egyptian Languages

According to Chen (1999), Ancient Egyptian language can be divided into five phases: Old Egyptian (c. 2700-2200 BCE), Middle Egyptian (c. 2200-1800 BCE), New Egyptian (c. 1580-700 BCE), Demotic (c. 700 BCE-600 CE), and Coptic (c. 600 CE-1000 CE). Since the Bronze Age lasted from approximately 3000 BCE to 700 BCE (Vandkilde 2016), only the Old Egyptian, Middle Egyptian, and New Egyptian phases of the Ancient Egyptian language were used during that time, while Demotic and Coptic phases came later. According to Loprieno and Müller (2012), Ancient Egyptian language represents a separate branch of the Afro-Asiatic language family.

8.3 Anatolian Languages

The historical region of Anatolia refers to the area that encompasses most of the territory of present-day Turkey. The Anatolian branch is the oldest attested branch of the Indo-European language family tree (Joseph 2003), but the languages belonging to the Anatolian branch are considered extinct (Pereltsvaig 2021). Anatolian languages include Hittite, Luwian, Palaic, Lycian, Milyan, Carian, Pisidian, and Sidetic (Adiego 2016).

Hittite language was written in cuneiform and used in areas that belong to present-day Turkey and northern Syria in the following periods: 1650-1500 BCE (old Hittite), 1500-1350 BCE (middle Hittite), and 1350-1180 BCE (new Hittite) (Sukhareva et al. 2017). Hittite is the oldest attested Indo-European language (Kloekhorst 2007). Grammatical comparison of Linear A and Hittite is available in Janke (2022).

The Luwian language was used in the central and southern Anatolia and northwestern Syria in the period c. 1500–700 BCE, and written in Anatolian hieroglyphs and an adaptation of Mesopotamian cuneiform (Yakubovich 2015).

Palaic language was used in the northern region of modern-day Turkey and attested from the sixteenth to twelfth century BCE (Bianconi 2019). Palaic language was written in cuneiform (Kudrinski 2018).

According to the attested timeframes given in Bianconi (2019) for the Carian (between the seventh and the third century BCE), Lycian[5] (between the sixth and the fourth century BCE), Sidetic (third century BCE), and Pisidian (between the first century BCE and the third century CE) languages, we can conclude that they were not used during the Bronze Age, whose approximate timeframe is given as 3000 BCE to 700 BCE in Vandkilde (2016).

One other language spoken in Anatolia during the Bronze Age, but not belonging to the Indo-European language family and Anatolian language branch, is Hattic. Hattic (or Hattian) language was used in Anatolia in the late third to mid-second millennium BCE, and the native speakers of the Hattic language did not record it in writing (Goedegebuure 2013). The only information we have on this language comes from the Hittite documents that occasionally included Hattic words and sentences (Revesz 2017b). According to Revesz (2017b), the Hattic language is generally considered by linguists to be a language isolate, but could possibly be an Ugric language and some form of Proto-Hattic could represent the origin of the Minoan language. According to the linguistic classification, the Ugric language is a branch of the Finno-Ugric language

5 Here we will also include the Milyan language (also known as Lycian B [Martínez Rodríguez 2021]), as it is a dialect of the Lycian language (Burgin 2010).

family (which might be a sub-branch of the larger Uralic language family) and encompasses Hungarian and Ob-Ugric sub-branches (Pereltsvaig 2021). The Hungarian sub-branch refers to the Hungarian language that is native to Hungary and represents one of the official languages of the European Union. The Ob-Ugric sub-branch encompasses Khanty and Mansi languages (Pereltsvaig 2021) that are spoken in Russia.

In Revesz (2017b) an expansion of the Uralic language family is proposed, where the Hungarian sub-branch is replaced with the West-Ugric sub-branch that encompasses Minoan, Hattic, and Hungarian languages. Revesz (2017b) also suggests that the Minoan language is additionally connected to both the Indo-European language family and the Greek language. In addition to Revesz (2017b), there have also been other papers connecting the Minoan language to the Finno-Ugric language branch (e.g., Schrijver 2019; Revesz 2018, 2020).

Finally, it is interesting to note that in late 2023 a new language was discovered on cuneiform tablets from Hattusa, Anatolia (Julius-Maximilians-Universität Würzburg 2023a,b; de Lazaro 2023; Keys 2023). According to de Lazaro (2023), it is believed that this newly discovered Indo-European language might have been spoken by the people from Kalasma (located approximately in modern north-western Turkey). Because of this, this new language has been named Kalasma, Kalašma, or Kalasmic language. Not much is currently known about this language, but it was approximately used from 1650 BCE to 1200 BCE (Julius-Maximilians-Universität Würzburg 2023a). The discovery of Kalasmic language gives hope to researchers working with ancient languages as it shows them that there are secrets waiting to be uncovered, even after thousands of years in waiting.

8.4 Phoenician Language

Phoenicia was an "ancient country of southwestern Asia at the eastern end of the Mediterranean Sea where modern Lebanon and adjacent parts of Syria and Israel now are" (Merriam-Webster.com Dictionary 2023b).

The Phoenician language was first spoken in the coastal parts of today's Lebanon from the twelfth century BCE until 196 CE, and spread to many Mediterranean areas by Phoenician merchants (Hetzron 2009). Linguistic classification of the Phoenician language is Afro-Asiatic $\rightarrow$ Proto-Semitic $\rightarrow$ West Semitic $\rightarrow$ Central Semitic $\rightarrow$ Northwest Semitic $\rightarrow$ Canaanite (Rubin 2008).

8.5 Later Languages

Most significant languages and/or language groups that became prominent after or at the brink of the disappearance of the Bronze Age Aegean and Cypriot scripts and in the relative vicinity of the Bronze Age Minoan civilization include the Etruscan language (central-west Italy and French island of Corsica), Messapic language (Southeastern Italy), Thracian language (Southeastern Europe), Illyrian language (Southeastern Europe), Phrygian language (Turkey), Dacian language (Southeastern Europe), Aramaic language (Middle East), Lepontic language (Northern Italy), and many other languages and dialects. Since these languages were used in the relative geographical proximity to Crete and Cyprus, they may have been influenced in some ways by the Minoan civilization and may hold the key to the decipherment of the Cretan hieroglyphic, Linear A, and Cypro-Minoan scripts.

People who spoke Etruscan and Messapic languages might have had possible connections to the Minoan civilization. The earliest known inscriptions in the Etruscan

language are dated to about 700 BCE (Freeman 1999). According to Robinson (2009), Herodotus wrote that the Etruscans migrated to Italy through the Aegean islands from Lydia in Anatolia, but most scholars disagree with that point of view due to the lack of archaeological evidence. Etruscan was spoken in central-west Italy, written in the Etruscan alphabet, and it still remains undeciphered. Messapic, on the other hand, was first attested in the sixth century BCE (Matzinger 2015), and was spoken in southeastern Italy. According to Blažek (2005), Messapian people are first mentioned by Herodotus as descendants of Cretans at the time of Minos. In Greek mythology Minos was the king of Crete associated with the labyrinth of Minotaur, and the king after which Sir Arthur J. Evans named the Minoan civilization.

9. Conclusion

In this article we presented a review of computational approaches to deciphering Bronze Age Aegean and Cypriot scripts, namely, the Archanes script and the Archanes formula, Cretan hieroglyphic (including the Malia Altar Stone and Arkalochori Axe), Phaistos Disk, Linear A, Linear B, Cypro-Minoan, and Cypriot scripts. We analyzed and compared the available digital corpora and resources associated with these scripts, outlined the possible challenges and proposed the required steps that need to be undertaken in order to improve the likelihood of computational decipherment of ancient scripts.

Our main conclusions are as follows:

- There is a dire need for a unified, digitalized dataset of Aegean and Cypriot inscriptions and dictionaries, alongside possible phonetic transcriptions of different symbols.
- This unified dataset should contain inscriptions and dictionaries regarding the other languages and scripts that might be related to Bronze Age Aegean and Cypriot scripts as well, mainly those used in ancient Anatolia, Mesopotamia, Phoenicia, and Egypt.
- Bronze Age Aegean and Cypriot inscriptions are scarce and short, so the augmentation of the proposed unified dataset should be considered, either computational or physical (i.e., via novel archaeological discoveries).
- More research (either genealogical, linguistic, or archaeological) is needed on the possible connections of Bronze Age Aegean and Cypriot scripts to other scripts and languages used in the relative geographical and chronological vicinity. This will probably require the use of high performance computing.
- And finally, more exposure on the currently undeciphered ancient scripts is needed. In our personal experience many people were unaware of the existence of these scripts, and the more information is released to the general public about them, the higher the probability of their successful decipherment becomes.

This article can serve as an introduction to some of the earliest European scripts and languages, and help shed some light onto the scripts and languages that came after them, such as the Illyrian language that is still not very well understood. In our future

work we plan on expanding our research to other undeciphered scripts and languages used geographically close to Crete and Cyprus (we will mainly focus on the Balkan Peninsula), but not necessarily in the same time period. We also plan on performing the linguistic typology analysis of Bronze Age Aegean and Cypriot languages, alongside their possibly related languages, in order to obtain more information on possible connections between these scripts. Furthermore, we are already working on a possible artificial intelligence and machine learning oriented analysis of low-resource languages, and plan on continuing and expanding our research in this increasingly important field.

Acknowledgments

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no 951732. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, United Kingdom, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Switzerland, Turkey, Republic of North Macedonia, Iceland, and Montenegro. This work was also partly supported by the Ministry of Science and Education of the Republic of Croatia under the FESB VIF project "iEnv - Intelligent Observers in Environmental Protection." The authors would like to thank Dr. Simon Castellan from INRIA Rennes-Bretagne Atlantique research centre for his swift and informative replies to their questions about the SigLA database, and would also like to express their gratitude to the anonymous reviewers who helped improve this article. Thank you.

References

Achterberg, Winfried. 2004. *The Phaistos Disc: A Luwian Letter to Nestor*, volume 13. Dutch Archaeological and Historical Society.

Adiego, Ignasi-Xavier. 2016. Anatolian languages and Proto-Indo-European. *Veleia*, 33:49–64. https://doi.org/10.1387/veleia.16819

Akhmetov, Iskander, Alexandr Pak, Irina Ualiyeva, and Alexander Gelbukh. 2020. Highly language-independent word lemmatization using a machine-learning classifier. *Computación y Sistemas*, 24(3):1353–1364. https://doi.org/10.13053/cys-24-3-3775

Allan, K. 2015. *The Routledge Handbook of Linguistics*. Routledge Handbooks in Linguistics. Taylor & Francis. https://doi.org/10.4324/9781315718453

Amazon. 2023. Amazon Web Services. https://aws.amazon.com/. [Online; accessed 21-April-2023].

Anastasiadou, Maria. 2016. Drawing the line: Seals, script, and regionalism in Protopalatial Crete. *American Journal of Archaeology*, 120(2):159–193. https://doi.org/10.3764/aja.120.2.0159

Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403*, pages 1–93.

Antony, P. J., Santhanu P. Mohan, and K. P. Soman. 2010. SVM based part of speech tagger for Malayalam. In *2010 International Conference on Recent Trends in Information, Telecommunication and Computing*, pages 339–341. https://doi.org/10.1109/ITC.2010.86

Assael, Yannis, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: A case study on Greek epigraphy. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6368–6375. https://doi.org/10.18653/v1/D19-1668

Assael, Yannis, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022a. ITHACA: Restoring and attributing ancient texts using deep neural networks. https://ithaca.deepmind.com/. [Online; accessed 8-January-2024].

Assael, Yannis, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022b. ITHACA: Restoring and attributing

ancient texts using deep neural networks. https://github.com/google-deepmind/ithaca. [Online; accessed 9-January -2024].
Assael, Yannis, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022c. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283. https://doi.org/10.1038/s41586-022-04448-z, PubMed: 35264762
Aurora, Federico. 2015a. DĀMOS (Database of Mycenaean at Oslo). Annotating a fragmentarily attested language. In *Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015), Procedia-Social and Behavioral Sciences*, volume 198, pages 21–31. https://doi.org/10.1016/j.sbspro.2015.07.415
Aurora, Federico. 2015b. DĀMOS (Database of Mycenaean at Oslo). Annotating a fragmentarily attested language. https://damos.hf.uio.no/about/online/. [Online; accessed 20-February-2023]. https://doi.org/10.1016/j.sbspro.2015.07.415
Azarine, Indira Suri, Moch Arif Bijaksana, and Ibnu Asror. 2019. Named entity recognition on Indonesian tweets using hidden Markov model. In *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pages 1–5. https://doi.org/10.1109/ICoICT.2019.8835277
Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, pages 1–15.
Beckman, Gary. 1983. Mesopotamians and Mesopotamian learning at Hattuša. *Journal of Cuneiform Studies*, 35(12):97–114. https://doi.org/10.2307/3515944
Bennett, Emmett L., John Chadwick, and Michael Ventris. 1956. The Knossos tablets: A revised transliteration of all the texts in Mycenaean Greek recoverable from Evans' excavations of 1900–1904. *Bulletin of the Institute of Classical Studies of the University of London. Supplementary Papers*, (2):1–125.
Berg-Kirkpatrick, Taylor and Dan Klein. 2011. Simple effective decipherment via combinatorial optimization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 313–321.
Bianconi, Michele. 2019. *The Linguistic Relationships between Greek and the Anatolian Languages*. Ph.D. thesis, University of Oxford. https://doi.org/10.1163/15699846-02001004
Billigmeier, Jon C. 1976. Toward a decipherment of Cypro-Minoan. *American Journal of Archaeology*, 80(3):295–300. https://doi.org/10.2307/503040
Blagec, Kathrin, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. 2022. A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 52–63. https://doi.org/10.18653/v1/2022.nlppower-1.6
Blažek, Václav. 2005. Paleo-Balkanian languages I: Hellenic languages. *Sborník Prací Filozofické Fakulty Brněnské Univerzity, Studia Minora Facultatis Philosophicae Universitatis Brunensis*, 54(10):15–33.
Bölücü, Necva and Burcu Can. 2019. Unsupervised joint POS tagging and stemming for agglutinative languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):1–21. https://doi.org/10.1145/3292398
Bontempi, Gianluca. 2021. *Statistical Foundations of Machine Learning: The Handbook*.
Brill, Eric. 1992. A simple rule-based part of speech tagger. In *Speech and Natural Language: Proceedings of a Workshop*, pages 152–155. https://doi.org/10.3115/974499.974526
Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
Burgin, James. 2010. A geographical note on the Xanthos stele. *Kadmos Bd. 49*, pages 181–186. https://doi.org/10.1515/kadmos.2010.011
Cahyani, Denis Eka and Mtchael Juan Vindiyanto. 2019. Indonesian part of speech tagging using hidden Markov model – Ngram & Viterbi. In *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*,

pages 353–358. https://doi.org/10.1109/ICITISEE48480.2019.9003989

Callaway, Ewen. 2013. Minoan civilization was made in Europe. *Nature*. https://doi.org/10.1038/nature.2013.12990

Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149. https://doi.org/10.1007/978-3-030-01264-9_9

Caron, Mathilde, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924.

Chadwick, John. 1987. *Linear B and Related Scripts*. 1. University of California Press.

Chadwick, John and Michael Ventris. 1973. *Documents in Mycenaean Greek*. Cambridge at the University Press, Second edition by John Chadwick.

Chen, Peter P. 1999. From Ancient Egyptian language to future conceptual modeling. In *Conceptual Modeling: Current Issues and Future Directions. Lecture Notes in Computer Science (LNCS)*, volume 1565. Springer, pages 56–64. https://doi.org/10.1007/3-540-48854-5_5

Chen, Xiaoyu, Shenghui Shi, Siyan Zhan, Daguang Jiang, and Xiaoyong Lin. 2019. Named entity recognition of Chinese electronic medical records based on cascaded conditional random field. In *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, pages 364–368. https://doi.org/10.1109/ICBDA.2019.8713244

Chiche, Alebachew and Betselot Yitagesu. 2022. Part of speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):1–25. https://doi.org/10.1186/s40537-022-00561-y

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. https://doi.org/10.3115/v1/D14-1179

Chorozoglou, G., N. Koukis, and E. C. Papakitsos. 2017. An application of software engineering for investigating the language of Phaistos disk. *Open Academic Journal of Advanced Science and Technology*, 1(1):20–29. https://doi.org/10.33094/5.2017.11.20.29

Civitillo, Matilde. 2021. R.E.A.D.I.N.G.: Cretan hieroglyphic inscriptions on seals. *Pasiphae*, XV:83–108.

Coe, Michael D. 1999. *Breaking the Maya Code*. Thames & Hudson.

Colin, Loh Jia Sheng and Francesco Perono Cacciafoco. 2020. A new approach to the decipherment of Linear A, stage 2. Cryptanalysis and language deciphering: A "brute force attack" on an undeciphered writing system. In Yannis Haralambous, editor, *Proceedings of Grapholinguistics in the 21st Century, Grapholinguistics and Its Applications*, volume 5. Brest: Fluxus Editions, pages 927–943.

Corazza, Michele, Silvia Ferrara, Barbara Montecchi, Fabio Tamburini, and Miguel Valério. 2021. The mathematical values of fraction signs in the Linear A script: A computational, statistical and typological approach. *Journal of Archaeological Science*, 125:1–14. https://doi.org/10.1016/j.jas.2020.105214

Corazza, Michele, Fabio Tamburini, Miguel Valério, and Silvia Ferrara. 2022. Unsupervised deep learning supports reclassification of Bronze Age Cypriot writing system. *PloS ONE*, 17(7):1–22. https://doi.org/10.1371/journal.pone.0269544, PubMed: 35834491

Costa-jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Élahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, pages 1–190.

Crane, Gregory. 2023a. Open Greek and Latin Perseus Digital Library Scaife Viewer. https://scaife.perseus.org/. [Online; accessed 9-January-2024].

Crane, Gregory. 2023b. The Perseus Digital Library and the future of libraries. *International Journal on Digital Libraries*, 24(2):117–128. https://doi.org/10.1007/s00799-022-00333-2

Croft, William. 2003. Typology. In Mark Aronoff and Janie Rees-Miller, editors, *The Handbook of Linguistics*. Blackwell

Publishers Ltd, Oxford, UK, chapter 14, pages 337–368. https://doi.org/10.1002/9780470756409.ch14

Cui, Leyang, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845. https://doi.org/10.18653/v1/2021.findings-acl.161

Currey, Anna and Alina Karakanta. 2016. Using related languages to enhance statistical language models. In *Proceedings of the NAACL Student Research Workshop*, pages 116–123. https://doi.org/10.18653/v1/N16-2017

Cushman, Ellen. 2011. The Cherokee syllabary: A writing system in its own right. *Written Communication*, 28:255–281. https://doi.org/10.1177/0741088311410172

Daggumati, Shruti and Peter Z. Revesz. 2019. Data mining ancient scripts to investigate their relationships and origins. In *Proceedings of the 23rd International Database Applications & Engineering Symposium*, pages 1–10. https://doi.org/10.1145/3331076.3331116

Daggumati, Shruti and Peter Z. Revesz. 2023. Convolutional neural networks analysis reveals three possible sources of Bronze Age writings between Greece and India. *Information*, 14(4):1–19. https://doi.org/10.3390/info14040227

Daniels, Peter and William Bright. 2010. *The World's Writing Systems*. Oxford University Press.

Daniels, Peter T. 2003. Writing systems. In *The Handbook of Linguistics*. John Wiley & Sons, Ltd, chapter 3, pages 43–80. https://doi.org/10.1002/9780470756409.ch3

Das, Avisha and Rakesh M. Verma. 2020. Can machines tell stories? A comparative study of deep neural language models and metrics. *IEEE Access*, 8:181258–181292. https://doi.org/10.1109/ACCESS.2020.3023421

Davis, Brent. 2010. Introduction to the Aegean pre-alphabetic scripts. *Kubaba*, 1:38–61. https://doi.org/10.31826/9781463233990-005

Davis, Brent. 2011. Cypro-Minoan in Philistia. *Kubaba*, 2:40–74.

Decorte, Roeland P.-J. E. 2018. The first 'European' writing: Redefining the Archanes Script. *Oxford Journal of Archaeology*, 37(4):341–372. https://doi.org/10.1111/ojoa.12152

de Lazaro, Enrico. 2023. 3,000-year-old cuneiform tablet reveals previously unknown language. Sci.News, https://www.sci.news/archaeology/kalasma-language-12294.html. [Online; accessed 8-December-2023].

Dereza, Oksana. 2018. Lemmatization for ancient languages: Rules or neural networks? In *Artificial Intelligence and Natural Language: 7th International Conference, AINL 2018*, pages 35–47. https://doi.org/10.1007/978-3-030-01204-5_4

Deri, Aliya and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408. https://doi.org/10.18653/v1/P16-1038

Ding, Chenchen, Masao Utiyama, and Eiichiro Sumita. 2018. Simplified abugidas. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 491–495. https://doi.org/10.18653/v1/P18-2078

Drovo, Mah Dian, Moithri Chowdhury, Saiful Islam Uday, and Amit Kumar Das. 2019. Named entity recognition in Bengali text using merged hidden Markov model and rule base approach. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. https://doi.org/10.1109/ICSCC.2019.8843661

Dryer, Matthew S. 1997. On the six-way word order typology. *Studies in Language*, 21(1):69–103. https://doi.org/10.1075/sl.21.1.04dry

Duhoux, Yves. 1990. Deciphering Bronze Age scripts of Crete - The case of Linear A. In *Advances in Cryptology — EUROCRYPT '89, LNCS 434*, pages 649–650. https://doi.org/10.1007/3-540-46885-4_61

Duhoux, Yves. 1998. Pre-Hellenic language(s) of Crete. *Journal of Indo-European Studies*, 26(1/2):1–39.

Eisenberg, Jerome M. 2008. The Phaistos Disk: A one hundred year old hoax. *Minerva*, 19:9–24.

Ekbal, Asif and Sivaji Bandyopadhyay. 2008. Part of speech tagging in Bengali using support vector machine. In *2008 International Conference on Information Technology*, pages 106–111. https://doi.org/10.1109/ICIT.2008.12

EMBL's European Bioinformatics Institute. 2023. ClustalW2. https://www.ebi

.ac.uk/Tools/msa/clustalw2/. [Online; accessed 21-December-2023].

Evans, Arthur J. 1909a. *Scripta Minoa: The written documents of Minoan Crete with Special Reference to the Archives of Knossos (Volume 2): The Hieroglyphic and Primitive Linear Classes*. Oxford at the Clarendon Press.

Evans, Arthur J. 1909b. *Scripta Minoa: The Written Documents of Minoan Crete with Special Reference to the Archives of Knossos (Volume I): The Hieroglyphic and Primitive Linear Classes*. Oxford at the Clarendon Press.

Ezhilarasi, S. and P. Uma Maheswari. 2021a. Designing the neural model for POS tag classification and prediction of words from ancient stone inscription script. *International Journal of Aquatic Science*, 12(3):1718–1728.

Ezhilarasi, S. and P. Uma Maheswari. 2021b. Depicting a neural model for lemmatization and POS tagging of words from palaeographic stone inscriptions. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1879–1884. https://doi.org/10.1109/ICICCS51141.2021.9432315

Facchetti, Giulio M. 2002. On some recent attempts to identify Linear A Minoan language. *Minos: Revista de Filología Egea*, 37:89–94.

Fernando, Sandareka, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. 2016. Comprehensive part-of-speech tag set and SVM based POS tagger for Sinhala. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 173–182.

Ferrara, Silvia, Kathryn Kelley, Barbara Montecchi, Ludovica Ottaviano, Miguel Valério, Andrea Santamaria, Mattia Cartolano, and Michele Corazza. 2023a. INSCRIBE - INvention of SCRIpts and their BEginnings. https://site.unibo.it/inscribe/en. [Online; accessed 13-June-2023].

Ferrara, Silvia, Kathryn Kelley, Barbara Montecchi, Ludovica Ottaviano, Miguel Valério, Andrea Santamaria, Mattia Cartolano, and Michele Corazza. 2023b. INSCRIBE 3D Interactive Web Viewer. https://www.inscribercproject.com/3d_viewer_home.php. [Online; accessed 20-February-2023].

Ferrara, Silvia, Barbara Montecchi, and Miguel Valério. 2021. Rationalizing the Cretan Hieroglyphic signlist. *Kadmos*, 60(1–2):5–32. https://doi.org/10.1515/kadmos-2021-0003

Ferrara, Silvia, Barbara Montecchi, and Miguel Valério. 2021. What is the 'Archanes formula'? Deconstructing and reconstructing the earliest attestation of writing in the Aegean. *Annual of the British School at Athens*, 116:43–62. https://doi.org/10.1017/S0068245420000155

Ferrara, Silvia, Barbara Montecchi, and Miguel Valério. 2022. The relationship between Cretan hieroglyphic and Linear A: A palaeographic and structural approach. *Pasiphae*, XVI:81–109.

Ferrara, Silvia, Barbara Montecchi, and Miguel Valério. 2023. In search of lost signs: A new approach to the issue of writing and non-writing on Cretan hieroglyphic seals. *Oxford Journal of Archaeology*, 42(2):107–130. https://doi.org/10.1111/ojoa.12265

Ferrara, Silvia and Fabio Tamburini. 2022. Advanced techniques for the decipherment of ancient scripts. *Lingue e Linguaggio*, 2/2022, July-December:239–259.

Ferrara, Silvia and Judith Weingarten. 2022. Cretan hieroglyphic seals and script: a view from the east. In Fabrizio Serra, editor, *Pasiphae: Rivista di Filologia e Antichitá Egee: XVI, 2022, Pisa*, pages 111–121.

Freeman, Philip. 1999. The survival of the Etruscan language. *Etruscan Studies*, 6(1):75–84. https://doi.org/10.1515/etst.1999.6.1.75

Freihat, Abed Alhakim, Mourad Abbas, Gábor Bella, and Fausto Giunchiglia. 2018. Towards an optimal solution to lemmatization in Arabic. *Procedia Computer Science*, 142:132–140. https://doi.org/10.1016/j.procs.2018.10.468

Frye, Richard N., Dietz O. Edzard, Wolfram Th. von Soden, The Editors of Encyclopaedia Britannica, Adam Augustyn, John Higgins, Gloria Lotha, J. E. Luebering, Marco Sampaolo, Shiveta Singh, Noah Tesch, Amy Tikkanen, Grace Young, and Adam Zeidan. 2023. History of Mesopotamia. Encyclopaedia Britannica, https://www.britannica.com/place/Mesopotamia-historical-region-Asia. [Online; accessed 19-May-2023].

Fuls, Andreas. 2015. Classifying undeciphered writing systems. *Historische Sprachforschung*, 128(1):42–58. https://doi.org/10.13109/hisp.2015.128.1.42

Gage, Philip. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Gao, Jianfeng and Mark Johnson. 2008. A comparison of Bayesian estimators for unsupervised hidden Markov model POS taggers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 344–352. https://doi.org/10.3115/1613715.1613761

Gelb, Ignace J. and Robert M. Whiting. 1975. Methods of decipherment. *Journal of the Royal Asiatic Society*, 107(2):95–104. https://doi.org/10.1017/S0035869X00132769

Giménez, Jesús and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 43–46.

Glassner, Jean Jacques. 2018. Writing in Elam. In Javier Álvarez Mon, Gian Pietro Basello, and Yasmina Wicks, editors, *The Elamite World*. Routledge, chapter 22. https://doi.org/10.4324/9781315658032-23

Glottolog. 2023. Languages. https://glottolog.org/glottolog/language. [Online; accessed 23-May-2023].

Gnanadesikan, Amalia E. 2009. *The Writing Revolution: Cuneiform to the Internet*. Wiley-Blackwell, Chichester, U.K. https://doi.org/10.1002/9781444304671

Godart, Louis and Jean-Pierre Olivier. 1985. Recueil des inscriptions en Linéaire A: Addenda, corrigenda, concordances, index et planches des signes, 5. *Etudes crétoises*, 21.

Goedegebuure, Petra. 2013. Hattic language. *The Encyclopedia of Ancient History*. https://doi.org/10.1002/9781444338386.wbeah24094

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning (Adaptive Computation and Machine Learning series)*. MIT Press.

Google. 2023. Google Cloud Platform. https://cloud.google.com/. [Online; accessed 21-April-2023].

Gordon, Cyrus H. 1997. Amorite and Eblaite. *The Semitic Languages*, pages 100–113.

Graham, Yvette. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137. https://doi.org/10.18653/v1/D15-1013

Greco, Alessandro, Georigia Flouda, and Erika Notti. 2023. The pa-i-to epigraphic project. https://www.paitoproject.it/en/home-2/. [Online; accessed 13-June-2023].

Gutherz, Gai, Shai Gordin, Luis Sáenz, Omer Levy, and Jonathan Berant. 2023. Translating Akkadian to English with neural machine translation. *PNAS Nexus*, 2(5):1–10. https://doi.org/10.1093/pnasnexus/pgad096, PubMed: 37143863

Haddow, Barry, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732. https://doi.org/10.1162/coli_a_00446

Hafeez, Rabab, Muhammad Waqas Anwar, Muhammad Hasan Jamal, Tayyaba Fatima, Julio César Martínez Espinosa, Luis Alonso Dzul López, Ernesto Bautista Thompson, and Imran Ashraf. 2023. Contextual Urdu lemmatization using recurrent neural network models. *Mathematics*, 11(2):1–20. https://doi.org/10.3390/math11020435

Hajiyeva, Galiba. 2019. The historical traces of ancient Sumerian language in dialect lexics of Azerbaijan and Turkish language. *International Journal of Innovative Technologies in Social Science*, 8(20):21–26. https://doi.org/10.31435/rsglobal_ijitss/30112019/6821

Hetzron, Robert. 2009. Semitic languages. Bernard Comrie, editor, *The World's Major Languages*, second edition. Routledge London, UK, chapter 32, pages 551–559.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Hofmanová, Zuzana, Susanne Kreutzer, Garrett Hellenthal, et al. 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences*, 113(25):6886–6891. https://doi.org/10.1073/pnas.1523951113, PubMed: 27274049

Hogan, Rob. 2022. Linear A and Linear B. https://github.com/mwenge. [Online; accessed 13-June-2023].

Hogan, Rob. 2023. Linear A Explorer. https://lineara.xyz/. [Online; accessed 4-September-2023].

Hsu, Chih-Wei and Chih-Jen Lin. 2002. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425.

https://doi.org/10.1109/72.991427, PubMed: 18244442

Hughey, Jeffery R., Peristera Paschou, Petros Drineas, Donald Mastropaolo, Dimitra M. Lotakis, Patrick A. Navas, Manolis Michalodimitrakis, John A. Stamatoyannopoulos, and George Stamatoyannopoulos. 2013. A European population in Minoan Bronze Age Crete. *Nature Communications*, 4:1–7. https://doi.org/10.1038/ncomms2871, PubMed: 23673646

Ingason, Anton Karl, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). In *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008*, pages 205–216. https://doi.org/10.1007/978-3-540-85287-2_20

Jabbar, Abdul, Sajid Iqbal, Manzoor Ilahi Tamimy, Shafiq Hussain, and Adnan Akhunzada. 2020. Empirical evaluation and study of text stemming algorithms. *Artificial Intelligence Review*, 53:5559–5588. https://doi.org/10.1007/s10462-020-09828-3

Janke, Richard Vallance. 2022. The influence of Hittite and digraphia on Minoan Linear A proto-Greek libation invocations, 34 pages. KONOSO Press.

Javaheripi, Mojan and Sébastien Bubeck. 2023. Phi-2: The surprising power of small language models. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/. [Online; accessed 8-January-2024].

Jin, Yanliang, Jinfei Xie, Weisi Guo, Can Luo, Dijia Wu, and Rui Wang. 2019. LSTM-CRF neural network with gated self attention for Chinese NER. *IEEE Access*, 7:136694–136703. https://doi.org/10.1109/ACCESS.2019.2942433

Joseph, Brian D. 2003. Evidentials: Summation, questions, prospects. In *Studies in Evidentiality (Typological Studies in Language)*, volume 54. John Benjamins Publishing Company, pages 307–327. https://doi.org/10.1075/tsl.54.17jos

Julius-Maximilians-Universität Würzburg. 2023a. New Indo-European language discovered. Julius-Maximilians-Universität Würzburg, https://www.uni-wuerzburg.de/en/news-and-events/news/detail/news/new-indo-european-language-discovered/. [Online; accessed 8-December-2023].

Julius-Maximilians-Universität Würzburg. 2023b. New Indo-European language discovered during excavation in Turkey. Phys.org, https://phys.org/news/2023-09-indo-european-language-excavation-turkey.html. [Online; accessed 8-December-2023].

Karajgikar, Jajwalya, Amira Al-Khulaidy, and Anamaria Berea. 2021. Computational pattern recognition in Linear A. *hal-03207615*. pages 1–18.

Karakanta, Alina, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32:167–189. https://doi.org/10.1007/s10590-017-9203-5

Kariyawasam, K. T. P. M., S. Y. Senanayake, and Prasanna S. Haddela. 2019. A rule based stemmer for Sinhala language. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 326–331. https://doi.org/10.1109/ICIIS47346.2019.9063286

Karnava, Artémis. (1999). *The Cretan Hieroglyphic Script of the Second Millennium BC: Description, Analysis, Function and Decipherment Perspectives (unpublished doctoral dissertation)*. Ph.D. thesis, Université libre de Bruxelles, Faculté de Philosophie et Lettres, Bruxelles.

Karnava, Artemis. 2014a. Cretan hieroglyphic script. In Georgios K. Giannakis, Vit Bubenik, Emilio Crespo, Chris Golston, Alexandra Lianeri, Silvia Luraghi, and Stephanos Matthaios, editors, *Encyclopedia of Ancient Greek Language and Linguistics, Volume 1, A–F*. Koninklijke Brill NV, pages 398–400.

Karnava, Artemis. 2014b. Cypriot syllabary. *Encyclopedia of Ancient Greek Language and Linguistics*. Brill, pages 404–408.

Karnava, Artemis. 2014c. Cypriot syllabary. In Georgios K. Giannakis, Vit Bubenik, Emilio Crespo, Chris Golston, Alexandra Lianeri, Silvia Luraghi, and Stephanos Matthaios, editors, *Encyclopedia of Ancient Greek Language and Linguistics, Volume 1, A–F*, volume 1. Koninklijke Brill NV, pages 404–408.

Karwatowski, Michal and Marcin Pietron. 2022. Context based lemmatizer for Polish language. *arXiv preprint arXiv:2207.11565*, pages 1–5.

Kaur, Harmanjeet and Preetpal Kaur Buttar. 2020. A rule-based stemmer for Punjabi adjectives. *International Journal of Advanced*

Research in Computer Science, 11(6):15–19. https://doi.org/10.26483/ijarcs.v11i6.6665

Kenanidis, Ioannis K. and Evangelos C. Papakitsos. 2015. A comparative linguistic study about the Sumerian influence on the creation of the Aegean scripts. *Scholars Journal of Arts, Humanities and Social Sciences*, 3(1E):332–346.

Kenanidis, Ioannis K. and Evangelos C. Papakitsos. 2017. An interpretation of the Malia stone inscription in terms of the Cretan Protolinear Script. *Terra Sebus. Acta Musei Sabesiensis*, 9:43–56.

Kestemont, Mike, Guy De Pauw, Renske van Nie, and Walter Daelemans. 2017. Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities*, 32(4):797–815. https://doi.org/10.1093/llc/fqw034

Keys, David. 2023. Archaeologists discover previously unknown ancient language. Independent, https://www.independent.co.uk/news/science/archaeology/hittite-ancient-language-turkey-ankara-b2451364.html. [Online; accessed 8-December-2023].

Kim, Young Bum and Benjamin Snyder. 2013. Unsupervised consonant-vowel prediction over hundreds of languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1527–1536.

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa, and Connie J. Mulligan. 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proceedings of the Royal Society B: Biological Sciences*, 276(1668):2703–2710. https://doi.org/10.1098/rspb.2009.0408, PubMed: 19403539

Kloekhorst, Alwin. 2007. *Etymological Dictionary of the Hittite Inherited Lexicon*. Brill, Leiden, The Netherlands.

Knight, Kevin, Anish Nair, Nishit Rathod, and Kenji Yamada. 2006. Unsupervised analysis for decipherment problems. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 499–506. https://doi.org/10.3115/1273073.1273138

Knight, Kevin and Kenji Yamada. 1999. A computational approach to deciphering unknown scripts. In *Unsupervised Learning in Natural Language Processing*, pages 37–44.

Kober, Alice E. 1948. The Minoan scripts: Fact and theory. *American Journal of Archaeology*, 52(1):82–103. https://doi.org/10.2307/500554

Kolinsky, Régine and Arlette Verhaeghe. 2017. Lace your mind: The impact of an extra-curricular activity on enantiomorphy. *Journal of Cultural Cognitive Science volume*, 1:57–64. https://doi.org/10.1007/s41809-017-0007-1

Krishnapriya, V., P. Sreesha, T. R. Harithalakshmi, T. C. Archana, and Jayasree N. Vettath. 2014. Design of a POS tagger using conditional random fields for Malayalam. In *2014 First International Conference on Computational Systems and Communications (ICCSC)*, pages 370–373. https://doi.org/10.1109/COMPSC.2014.7032680

Kudo, Taku. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. https://doi.org/10.18653/v1/P18-1007

Kudrinski, Maksim. 2016. Hittite heterographic writings and their interpretation. *Indogermanische Forschungen*, 121(1):159–176. https://doi.org/10.1515/if-2016-0009

Kudrinski, Maksim. 2018. Heterograms in Hittite, Palaic, and Luwian context. *Journal of Language Relationship*, 15(3–4):238–249. https://doi.org/10.31826/9781463239909-009

Lafferty, John D., Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. https://doi.org/10.18653/v1/N16-1030

Lastilla, Lorenzo, Roberta Ravanelli, and Silvia Ferrara. 2019. 3d high-quality modeling of small and complex archaeological inscribed objects: Relevant issues and proposed methodology. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, GEORES 2019 - 2nd International*

Conference of Geomatics and Restoration, XLII-2/W11:699–706. https://doi.org/10.5194/isprs-archives-XLII-2-W11-699-2019

Lazaridis, Iosif, Alissa Mittnik, Nick Patterson, Swapan Mallick, Nadin Rohland, Saskia Pfrengle, Anja Furtwängler, Alexander Peltzer, Cosimo Posth, Andonis Vasilakis, et al. 2017. Genetic origins of the Minoans and Mycenaeans. *Nature*, 548(7666):214–218. https://doi.org/10.1038/nature23310, PubMed: 28783727

Lee, Sang Zoo, Jun'ichi Tsujii, and Hae Chang Rim. 2000. Part-of-speech tagging based on hidden Markov model assuming joint independence. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 263–269. https://doi.org/10.3115/1075218.1075252

Lee, Seungjun, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1–22. https://doi.org/10.3390/math11041006

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Lin, Chin Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Loprieno, Antonio and Matthias Müller. 2012. Ancient Egyptian and Coptic. In Zygmunt Frajzyngier and Erin Shay, editors, *The Afroasiatic Languages*. Cambridge University Press, chapter 3, pages 102–144.

Luo, Jiaming, Yuan Cao, and Regina Barzilay. 2019a. 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. https://doi.org/10.18653/v1/P19-1303

Luo, Jiaming, Yuan Cao, and Regina Barzilay. 2019b. Neural decipherment via minimum-cost flow: From Ugaritic to Linear B. https://github.com/j-luo93/NeuroDecipher. [Online; accessed 5-September-2023].

Luo, Jiaming, Frederik Hartmann, Enrico Santus, Regina Barzilay, and Yuan Cao. 2021. Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the Association for Computational Linguistics*, 9:69–81. https://doi.org/10.1162/tacl_a_00354

Manjavacas, Enrique, Akos Kadar, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1493–1503. https://doi.org/10.18653/v1/N19-1153

Martínez Rodríguez, Elena. 2021. Milyan dialectal divergence and its traces in onomastics. *Kadmos*, 60(1–2):137–156. https://doi.org/10.1515/kadmos-2021-0010

Matzinger, Joachin. 2015. Messapico e illirico. *L'Idomeneo*, 2015(19):57–66.

Mavridaki, Argyro, Eleni Galiotou, and Evangelos Papakitsos. 2021a. Designing a software application for the multilingual processing of the Linear A script. In *24th Pan-Hellenic Conference on Informatics*, PCI 2020, pages 167–169. https://doi.org/10.1145/3437120.3437299

Mavridaki, Argyro, Eleni Galiotou, and Evangelos C. Papakitsos. 2021b. Developing a software application for the study and learning of Linear A script. *Review of Computer Engineering Research*, 8(1):8–13. https://doi.org/10.18488/journal.76.2021.81.8.13

Melena, José L. 2014. Mycenaean writing. In Yves Duhoux and Anna Morpurgo Davies, editors, *A Companion to Linear B: Mycenaean Greek Texts and Their World*, volume 3, Peeters Louvain-la-Neuve – Walpole, MA, chapter 17.

Melucci, Massimo and Nicola Orio. 2003. A novel method for stemmer generation based on hidden Markov models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 131–138. https://doi.org/10.1145/956863.956889

Merriam-Webster.com Dictionary, Merriam-Webster. 2023a. Boustrophedon. https://www.merriam-webster.com

/dictionary/boustrophedon. [Online; accessed 28-February-2023].
Merriam-Webster.com Dictionary, Merriam Webster. 2023b. Phoenicia. https://www.merriam-webster.com/dictionary/Phoenicia. [Online; accessed 23-May-2023].
Microsoft. 2023. Microsoft Azure. https://azure.microsoft.com/en-us. [Online; accessed 21-April-2023].
Mielke, Sabrina J., Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, et al. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint arXiv:2112.10508*, pages 1–27.
Min Eu, Niki Cassandra, Duo Duo Xu, and Francesco Perono Cacciafoco. 2019. Coding to decipher Linear A. In *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, pages 1–6.
Mittal, Sumeer, Navdeep Singh Sethi, and Sanjeev Kumar Sharma. 2014. Part of speech tagging of Punjabi language using N gram model. *International Journal of Computer Applications*, 100(19):19–23. https://doi.org/10.5120/17634-8229
Moran, Steven and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.
Moran, Steven, Daniel McCloy, and Richard Wright, editors. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
Mycenaean Epigraphy Group, University of Cambridge, Faculty of Classics. 2023. The decipherment of Linear B: Introduction – The decipherment process. https://www.classics.cam.ac.uk/system/files/documents/process.pdf. [Online; accessed 31-August-2023].
Nakagawa, Tetsuji, Taku Kudo, and Yuji Matsumoto. 2001. Unknown word guessing and part-of-speech tagging using support vector machines. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS)*, pages 325–331.
Nandathilaka, Maheshi, Supunmali Ahangama, and G. Thilini Weerasuriya. 2018. A rule-based lemmatizing approach for Sinhala language. In *2018 3rd International Conference on Information Technology Research (ICITR)*, pages 1–5. https://doi.org/10.1109/ICITR.2018.8736134
Neef, Martin. 2015. Writing systems as modular objects: Proposals for theory design in grapholinguistics. *Open Linguistics*, 1. https://doi.org/10.1515/opli-2015-0026
Nosch, Marie Louise and Agata Ulanowska. 2021. The materiality of the Cretan Hieroglyphic script: Textile production-related referents to hieroglyphic signs on seals and sealings from Middle Bronze Age Crete. In Philip John Boyes, Philippa M. Steele, and Natalia Elvira Astoreca, editors, *The Social and Cultural Contexts of Historic Writing Practices*, volume 2. Oxbow Books, chapter 5, pages 73–100. https://doi.org/10.2307/j.ctv2npq9fw.10
Oakes, Michael Philip. 2019. Statistical analysis of the tables in Mahadevan's concordance of the Indus Valley script. *Journal of Quantitative Linguistics*, 26(1):22–47. https://doi.org/10.1080/09296174.2017.1406294
Olivier, Jean Pierre, Louis Godart, and Jean-Claude Poursat. 1996. Corpus Hieroglyphicarum Inscriptionum Cretae. *Études Crétoises*, 31:1–447.
Omrak, Ayça, Torsten Günther, Cristina Valdiosera, Emma M. Svensson, Helena Malmström, Henrike Kiesewetter, William Aylwardz, Jan Storå, Mattias Jakobsson, and Anders Götherström. 2016. Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. *Current Biology*, 26(2):270–275. https://doi.org/10.1016/j.cub.2015.12.019, PubMed: 26748850
OpenAI. 2023. ChatGPT. https://chat.openai.com/. [Online; accessed 8-January-2024].
Osborne, Margaret. 2022. Scientists translate the oldest sentence written in the first alphabet. https://www:smithsonianmag:com/smartnews/scientists-translate-the-oldest-sentence-written-in-thefirst-alphabet-180981101/. [Online; accessed 28-February-2023].
Owens, Gareth A. 1996. The common origin of Cretan hieroglyphs and Linear A. *Kadmos Bd.*, 35(2):105–110. https://doi.org/10.1515/kadm.1996.35.2.105
Pae, Hye K. and Min Wang. 2022. The effects of writing systems and scripts on cognition and beyond: An introduction. *Reading and Writing*, 35:1315–1321. https://doi.org/10.1007/s11145-022-10289-z
Pagel, Mark. 2017. Q&A: What is human language, when did it evolve and why should we care? *BMC Biology*, 15(64):1–6.

https://doi.org/10.1186/s12915-017-0405-3, PubMed: 28738867

Pallavi, A. S. P. and A. S. Pillai. 2014. Parts of speech (POS) tagger for Kannada using conditional random fields (CRFs). In *Proceedings of the National Conference on Indian Language Computing, NCILC*, pages 1–4 (4 pages).

Papakitsos, Evangelos C. and Ioannis K. Kenanidis. 2015. Additional palaeographic evidence for the relationship of the Aegean scripts to the Sumerian pictography. *Scholars Journal of Arts, Humanities and Social Sciences*, 3(3C):734–737.

Papavassileiou, Katerina, Gareth Owens, and Dimitrios Kosmopoulos. 2020. A dataset of Mycenaean Linear B sequences. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2552–2561.

Papavassileiou, Katerina, Dimitrios I. Kosmopoulos, and Gareth Owens. 2023. A generative model for the Mycenaean Linear B script and its application in infilling text from ancient tablets. *ACM Journal on Computing and Cultural Heritage*, 16(3):1–25. https://doi.org/10.1145/3593431

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. https://doi.org/10.3115/1073083.1073135

Patil, Nita, Ajay Patil, and B. V. Pawar. 2020. Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188. https://doi.org/10.1016/j.procs.2020.03.431

Pearce, Kristin M. 2010. The adaption of Akkadian into Cuneiform. *Colonial Academic Alliance Undergraduate Research Journal*, 1(1):1–10.

Pereltsvaig, Asya. 2021. *Languages of the World*, third edition. Cambridge University Press.

Petrolito, Tommaso, Ruggero Petrolito, Francesco Perono Cacciafoco, and Grégoire Winterstein. 2015. Minoan linguistic resources: The Linear A digital corpus. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 95–104. https://doi.org/10.18653/v1/W15-3715

Plisson, Joël, Nada Lavrac, Dunja Mladenic, et al. 2004. A rule based approach to word lemmatization. In *Proceedings of IS*, volume 3, pages 83–86.

Porter, Martin F. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137. https://doi.org/10.1108/eb046814

Pourdamghani, Nima and Kevin Knight. 2017. Deciphering related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2513–2518. https://doi.org/10.18653/v1/D17-1266

Pourdamghani, Nima and Kevin Knight. 2019. Neighbors helping the poor: Improving low-resource machine translation using related languages. *Machine Translation*, 33(3):239–258. https://doi.org/10.1007/s10590-019-09236-7

Pramana, Rio, Debora, Jonathan Jansen Subroto, Alexander Agung Santoso Gunawan, and Anderies. 2022. Systematic literature review of stemming and lemmatization performance for sentence similarity. In *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, pages 1–6. https://doi.org/10.1109/ICITDA55840.2022.9971451

QGIS Development Team. 2023. *QGIS Geographic Information System*. QGIS Association.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. Technical report, pages 1–12.

Radikov, I., M. Pitukhina, O. Tolstoguzov, and V. Volokh. 2019. Experience of observation of indigenous minorities and ethnic minorities of Karelia. In *IOP Conference Series: Earth and Environmental Science*, volume 302(1), pages 1–8. https://doi.org/10.1088/1755-1315/302/1/012080

Rashid, M. Pervez. 2019. The design and implementation of AIDA: Ancient Inscription Database and Analytics system. Master's thesis, University of Nebraska - Lincoln.

Rau, Jeremy. 2010. Greek and Proto-Indo-European. In Egbert J. Bakker, editor, *A Companion to the Ancient Greek Language*. Wiley-Blackwell, chapter 12, pages 171–188. https://doi.org/10.1002/9781444317398.ch12

Reczko, Wolfgang. 2009. Analyzing and dating the structure of the Phaistos Disk. *Archaeological and Anthropological Sciences*, 1:241–245. https://doi.org/10.1007/s12520-009-0015-2

Remondino, Fabio, Stefano Girardi, Lorenzo Gonzo, and Franco Nicolis. 2008. Detailed 3D reconstruction of the great inscription of Gortyna, Crete: Acquisition, registration and visualization of multi-resolution data. In *Digital Heritage - Proceedings of 14th International Conference on Virtual Systems and MultiMedia (VSMM 2008)*, pages 404–412.

Revesz, Peter Z. 2015a. A computational study of the evolution of Cretan and related scripts. In *Mathematical Models and Computational Methods (Joint Proceedings of AMCSE-MMMAS-EAS), INASE Press*, pages 101–105.

Revesz, Peter Z. 2015b. A computational translation of the Phaistos Disk. *Mathematical Models and Computational Methods*, pages 53–57.

Revesz, Peter Z. 2016a. Bioinformatics evolutionary tree algorithms reveal the history of the Cretan Script Family. *International Journal of Applied Mathematics and Informatics*, 10(1):67–76.

Revesz, Peter Z. 2016b. A computer-aided translation of the Cretan Hieroglyph script. *International Journal of Signal Processing*, 1:127–133.

Revesz, Peter Z. 2016c. A computer-aided translation of the Phaistos Disk. *International Journal of Computers*, 10:94–100.

Revesz, Peter Z. 2017a. The Cretan script family includes the Carian Alphabet. In *MATEC Web of Conferences, 21st International Conference on Circuits, Systems, Communications and Computers (CSCC 2017), 05019*, volume 125, pages 1–4. https://doi.org/10.1051/matecconf/201712505019

Revesz, Peter Z. 2017b. Establishing the West-Ugric language family with Minoan, Hattic and Hungarian by a decipherment of Linear A. *WSEAS Transactions on Information Science and Applications*, 14(1):306–335.

Revesz, Peter Z. 2017c. A translation of the Arkalochori Axe and the Malia Altar Stone. *WSEAS Transactions on Information Science and Applications*, 14(1):124–133.

Revesz, Peter Z. 2017d. A translation of the Malia Altar Stone. *MATEC Web of Conferences 125, 05018, 21st International Conference on Circuits, Systems, Communications and Computers (CSCC 2017)*, 125:1–5. https://doi.org/10.1051/matecconf/201712505018

Revesz, Peter Z. 2018. Computational linguistics techniques for the study of ancient languages. *MATEC Web of Conferences 210, 03014, 22nd International Conference on Circuits, Systems, Communications and Computers (CSCC 2018)*, pages 1–5. https://doi.org/10.1051/matecconf/201821003014

Revesz, Peter Z. 2019. Minoan archaeogenetic data mining reveals Danube Basin and western Black Sea littoral origin. *International Journal of Biology and Biomedical Engineering*, 13:108–120.

Revesz, Peter Z. 2020. Minoan and Finno-Ugric regular sound changes discovered by data mining. In *2020 24th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, pages 241–246. https://doi.org/10.1109/CSCC49995.2020.00051

Revesz, Peter Z. 2021. Data mining autosomal archaeogenetic data to determine Minoan origins. In *Proceedings of the 25th International Database Engineering & Applications Symposium*, pages 46–55. https://doi.org/10.1145/3472163.3472178

Revesz, Peter Z. 2022. Experimental evidence for a left-to-right reading direction of the Phaistos Disk. *Mediterranean Archaeology and Archaeometry*, 22(1):79–96.

Revesz, Peter Z., M. Parvez Rashid, and Yves Tuyishime. 2019a. AIDA (Ancient Inscription Database and Analytics) system. https://cse.unl.edu/~revesz/index7.php. [Online; accessed 13-June-2023].

Revesz, Peter Z., M. Parvez Rashid, and Yves Tuyishime. 2019b. The design and implementation of AIDA: Ancient Inscription Database and Analytics system. In *Proceedings of the 23rd International Database Applications & Engineering Symposium (IDEAS '19)*, pages 1–6. https://doi.org/10.1145/3331076.3331117

Riaz, Kashif. 2010. Rule-based named entity recognition in Urdu. In *Proceedings of the 2010 Named Entities Workshop*, pages 126–135.

Robertson, Paul. 1999. GRAVA-a corpus based approach to the interpretation of aerial images. In *Image Processing And Its Applications, 1999. Seventh International Conference on (Conf. Publ. No. 465)*, volume 2, pages 527–531. https://doi.org/10.1049/cp:19990378

Robertson, Paul. 2001. *A Self Adaptive Architecture for Image Understanding*. Ph.D. thesis, University of Oxford.

Robinson, Andrew. 2007. *The Story of Writing: Alphabets, Hieroglyphs & Pictograms*. Thames & Hudson.

Robinson, Andrew. 2009. *Lost Languages: The Enigma of the World's Undeciphered Scripts*. Thames & Hudson.

Rubin, Aaron D. 2008. The subgrouping of the Semitic languages. *Language and Linguistics Compass*, 2(1):61–84. https://doi.org/10.1111/j.1749-818X.2007.00044.x

Sahala, Aleksi, Tero Alstola, Jonathan Valk, and Krister Lindén. 2023. Lemmatizing and POS-tagging Akkadian with BabyLemmatizer and dictionary-based post-correction. In *Selected Papers from the CLARIN Annual Conference 2022*, pages 111–119. https://doi.org/10.3384/ecp198011

Salgarella, Ester. 2020. Reconstruction of an orthographic system: The Linear B syllabary of Bronze Age Greece. In *Advances in Historical Orthography, c. 1500–1800*. Cambridge University Press.

Salgarella, Ester. 2022. Linear A. In *Oxford Classical Dictionary*. Oxford University Press.

Salgarella, Ester and Simon Castellan. 2021a. SigLA: The signs of Linear A: A palæographical database. *Grapholinguistics and Its Applications*, 5:945–962. https://doi.org/10.36824/2020-graf-salg

Salgarella, Ester and Simon Castellan. 2021b. SigLA: The signs of Linear A: A palæographical database. https://sigla.phis.me/. [Online; accessed 13-June-2023].

Saxe, Andrew, Stephanie Nelli, and Christopher Summerfield. 2021. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67. https://doi.org/10.1038/s41583-020-00395-8, PubMed: 33199854

Schoep, Ilse. 2002. The administration of neopalatial Crete: A critical assessment of the Linear A tablets and their role in the administrative process. *Minos: Revista de Filología Egea*, 17:1–230.

Schrijver, Peter. 2019. Talking Neolithic: The case for Hatto-Minoan and its relation to Sumerian. In Guus Kroonen, James P. Mallory, and Bernard Comrie, editors, *Talking Neolithic: Proceedings of the Workshop on Indo-European Origins Monograph No. 65*. Journal of Indo-European Studies, pages 336–374.

Schuster, Mike and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. https://doi.org/10.1109/ICASSP.2012.6289079

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162

Serafimov, Pavel and Giancarlo Tomezzoli. 2011. Evidence for early Slavic presence in Minoan Crete. In *Proceedings of the 9th International Topical Conference Origin of Europeans*, pages 219–229.

Seri, Andrea. 2010. Adaptation of cuneiform to write Akkadian. In Christopher Woods, editor, *Visible Language. Inventions of Writing in the Ancient Middle East and Beyond*, volume 32. The University of Chicago, chapter 3, pages 85–98.

Siewierska, Anna and Ludmila Uhlířová. 1998. An overview of word order in Slavic languages. In Anna Siewierska, editor, *Constituent Order in the Languages of Europe*. De Gruyter Mouton, Berlin, New York, pages 105–149. https://doi.org/10.1515/9783110812206.105

Šincek, Marijana. 2020. On, ona, ono: Translating gender neutral pronouns into Croatian. *Zbornik Radova Međunarodnog Simpozija Mladih Anglista, Kroatista i Talijanista*, pages 92–112.

Singh, Jasmeet and Vishal Gupta. 2017. An efficient corpus-based stemmer. *Cognitive Computation*, 9:671–688. https://doi.org/10.1007/s12559-017-9479-z

Singh, Jasmeet and Vishal Gupta. 2019. A novel unsupervised corpus-based stemming technique using lexicon and corpus statistics. *Knowledge-Based Systems*, 180:147–162. https://doi.org/10.1016/j.knosys.2019.05.025

Skelton, Christina. 2008. Methods of using phylogenetic systematics to reconstruct the history of the Linear B script. *Archaeometry*, 50(1):158–176. https://doi.org/10.1111/j.1475-4754.2007.00349.x

Smith, Joanna S. and Nicolle E. Hirschfeld. 1999. The Cypro-Minoan corpus project takes an archaeological approach. *Near Eastern Archaeology*, 62(2):129–130. https://doi.org/10.2307/3210706

Snyder, Benjamin. 2010. *Unsupervised Multilingual Learning*. Ph.D. thesis, Massachusetts Institute of Technology.

Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745. https://doi.org/10.3115/1613715.1613851

Snyder, Benjamin and Regina Barzilay. 2010. Climbing the tower of Babel: Unsupervised multilingual learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 29–36, Omnipress, Haifa, Israel.

Snyder, Benjamin, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1048–1057.

Sommerschield, Thea, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, pages 1–44. https://doi.org/10.1162/coli_a_00481

Spathis, Dimitris and Fahim Kawsar. 2023. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *arXiv preprint arXiv:2309.06236*. pages 1–5.

Sproat, Richard. 2014. A statistical comparison of written language and nonlinguistic symbol systems. *Language*, 90(2):457–481. https://doi.org/10.1353/lan.2014.0031

Sproat, Richard and Alexander Gutkin. 2021. The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, 47(3):477–528. https://doi.org/10.1162/coli_a_00409

Srivatsan, Nikita, Jason Vega, Christina Skelton, and Taylor Berg-Kirkpatrick. 2021. Neural representation learning for scribal hands of Linear B. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Proceedings, Part II 16*, pages 325–338. https://doi.org/10.1007/978-3-030-86159-9_23

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 507–514.

Stratos, Karl, Michael Collins, and Daniel Hsu. 2016. Unsupervised part-of-speech tagging with anchor hidden Markov models. *Transactions of the Association for Computational Linguistics*, 4:245–257. https://doi.org/10.1162/tacl_a_00096

Studiawan, Hudan, Mhd Fadly Hasan, and Baskoro Adi Pratomo. 2023. Rule-based entity recognition for forensic timeline. In *2023 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6. https://doi.org/10.1109/ICTAS56421.2023.10082742

Sukhareva, Maria, Francesco Fuscagni, Johannes Daxenberger, Susanne Görke, Doris Prechel, and Iryna Gurevych. 2017. Distantly supervised POS tagging of low-resource languages under extreme data sparsity: The case of Hittite. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 95–104. https://doi.org/10.18653/v1/W17-2213

Tan, Kimberly Miracle Wei Yan. 2022. Understanding Linear A through the lens of maritime history during the Bronze Age. Master's thesis, Nanyang Technological University.

Terras, Melissa. 2006a. *Image to Interpretation: An Intelligent System to Aid Historians in Reading the Vindolanda Texts*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199204557.001.0001

Terras, Melissa. 2006b. Interpreting the image: Using advanced computational techniques to read the Vindolanda texts. In *Aslib Proceedings*, volume 58(1/2), pages 102–117. https://doi.org/10.1108/00012530610648707

Terras, Melissa and Paul Robertson. 2005. Image and interpretation using artificial intelligence to read ancient roman texts. *Human IT*, 7(3):1–56.

Tomas, Helena. 2010. Linear A versus Linear B administrative systems in the sphere of religious matters. *MOM Éditions*, 54(1):121–133.

Tóth, Alfréd. 2007. Are all agglutinative languages related to one another? *Mikes International*, pages 1–28.

Tselentis, Chris. 2011. Linear B Lexicon. https://archive.org/details/LinearBLexicon/page/n5/mode/2up. [Online; accessed 5-September-2023].

Urban, Matthias. 2021. The geography and development of language isolates. *Royal Society Open Science*, 8(4):1–17. https://doi.org/10.1098/rsos.202232, PubMed: 33996125

Vainstub, Daniel, Madeleine Mumcuoglu, Michael G. Hasel, Katherine M. Hesler, Miriam Lavi, Rivka Rabinovich, Yuval Goren, and Yosef Garfinkel. 2022. A Canaanite's wish to eradicate lice on an inscribed ivory comb from Lachish. *Jerusalem Journal of Archaeology*, 2(2021–2022):76–119. https://doi.org/10.52486/01.00002.4

Valério, Miguel Filipe Grandão. 2016. *Investigating the signs and sounds of Cypro-Minoan*. Ph.D. thesis, Universitat de Barcelona.

Vandkilde, Helle. 2016. Bronzization: The Bronze Age as pre-modern globalization. *Praehistorische Zeitschrift*, 91(1):103–123. https://doi.org/10.1515/pz-2016-0005

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:2–11.

Vatri, Alessandro and Barbara McGillivray. 2020. Lemmatization for Ancient Greek: An experimental assessment of the state of the art. *Journal of Greek Linguistics*, 20(2):179–196. https://doi.org/10.1163/15699846-02002001

Ventris, Michael and John Chadwick. 2015. *Documents in Mycenaean Greek: Three Hundred Selected Tablets from Knossos, Pylos and Mycenae with Commentary and Vocabulary*. Cambridge University Press, Reprint edition.

Vidal-Gorène, Chahan and Bastien Kindt. 2020. Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old Georgian, and Syriac. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 22–27.

Viterbi, Andrew J. 2006. A personal history of the Viterbi algorithm. *IEEE Signal Processing Magazine*, 23(4):120–142. https://doi.org/10.1109/MSP.2006.1657823

Whittaker, Heléne. 2005. Social and symbolic aspects of Minoan writing. *European Journal of Archaeology*, 8(1):29–41. https://doi.org/10.1177/1461957105058207

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, pages 1–23.

Yakubovich, Ilya. 2015. The Luwian language. In *Oxford Handbook Topics in Linguistics*. Oxford. https://doi.org/10.1093/oxfordhb/9780199935345.013.18

Yan, Rongen, Xue Jiang, and Depeng Dang. 2021. Named entity recognition by using XLNet-BiLSTM-CRF. *Neural Processing Letters*, 53(5):3339–3356. https://doi.org/10.1007/s11063-021-10547-1

Yi, Feng, Bo Jiang, Lu Wang, and Jianjun Wu. 2020. Cybersecurity named entity recognition using multi-modal ensemble learning. *IEEE Access*, 8:63214–63224. https://doi.org/10.1109/ACCESS.2020.2984582

Younger, John G. 1999. The Cretan Hieroglyphic script: A review article. *Minos*, 31–32:379–400.

Younger, John G. 2023. John G. Younger. https://www.people.ku.edu/~jyounger/. [Online; accessed 13-June-2023].

Zampieri, Marcos, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612. https://doi.org/10.1017/S1351324920000492

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, pages 1–43.

Zin, Khine Khine and Ni Lar Thein. 2009. Part of speech tagging for Myanmar using hidden Markov model. In *2009 International Conference on the Current Trends in Information Technology (CTIT)*, pages 1–6. https://doi.org/10.1109/CTIT.2009.5423133