

A Dataset of Mycenaean Linear B Sequences

K. Papavasileiou¹, G. Owens², D. Kosmopoulos³

^{1,3}University of Patras
GR-26504, Rio

{¹cpapavas, ³dkosmo}@upatras.gr

²Hellenic Mediterranean University
GR-71410, Heraklion

ogareth@hmu.gr

Abstract

We present our work towards a dataset of Mycenaean Linear B sequences gathered from the Mycenaean inscriptions written in the 13th and 14th century B.C. (c. 1400-1200 B.C.). The dataset contains sequences of Mycenaean words and ideograms according to the rules of the Mycenaean Greek language in the Late Bronze Age. Our ultimate goal is to contribute to the study, reading and understanding of ancient scripts and languages. Focusing on sequences, we seek to exploit the structure of the entire language, not just the Mycenaean vocabulary, to analyse sequential patterns. We use the dataset to experiment on estimating the missing symbols in damaged inscriptions.

Keywords: Mycenaean Linear B script, sequential patterns

1. Introduction

The Mycenaean script constitutes one of the writing systems used in the Aegean in the 2nd millennium B.C. Mycenaean Linear B, was a syllabic script deciphered by the English architect Michael Ventris in 1-6-1952, who proved that it expresses an archaic form of the Greek language. The Mycenaean Linear B texts, though they are brief and terse revealing little about structure, provide valuable information about unknown aspects of Mycenaean life.

Despite the huge importance of the archaeological data, the written sources are those that highlight aspects of the political and social organization of the cultures under study, supporting the archaeological evidence with “historically” substantiated elements. For example, the decipherment of Mycenaean Linear B script added seven centuries to the history of the Hellenic language. Until then the oldest texts for the history of the Greek language were considered to be the Homeric epics.

The Greek language is among the oldest languages and therefore a valuable tool for linguistic observation (Ruiperez & Melena, 1996). However, the importance of Mycenaean documents is remarkable, not only for linguistics and philology, but also for other sciences such as religious studies, ethnology, history and law, since from the Mycenaean texts is obtained information on the political and administrative organization, the social structure, the economic activity, the religion and the military aspect of the Mycenaean civilization.

The computational methods are expected to offer a lot in decipherment tasks. Computational techniques (such as smoothed n-grams, Hidden Markov Models, Bayesian classifiers, Conditional Random Fields, etc.) might be applied to the problem of decipherment of ancient scripts (Knight & Yamada, 1999; Knight, et al., 2006; Ravi & Knight, 2008; Ravi & Knight, 2011b; Snyder, Barzilay, & Knight, 2010; Corlett & Penn, 2010; Nuhn, Mauser, & Ney, 2012; Nuhn & Ney, 2013) and also to the in-depth linguistic analysis (e.g., sentence detection, tokenization, lemmatization, part-of-speech tagging, etc.) of the already deciphered ones. The goal is not to replace the experts and their insight, but to contribute to their efforts by facilitating computational intelligence perspectives.

By presenting a dataset of Mycenaean sequences we aim to contribute to the restoration of the words of the damaged Mycenaean inscriptions. In the future we aspire to complement the decipherment efforts of the Minoan script. To this end we employ probabilistic methods for

structure prediction, in sequences such as the Conditional Random Fields (CRF), which are applicable in many areas including natural language processing, computer vision and bioinformatics (Sutton & McCallum, 2012).

In the next section we present the Mycenaean Linear B datasets used so far and the importance of our contribution. In section 3 we analyse the methods used to construct the dataset. In section 4 we present the resulting dataset. Section 5 describes an initial experiment on predicting missing symbols and discusses the results. Finally, section 6 concludes this work.

2. State of the art and contribution

After the decipherment of the Mycenaean Linear B script there was an immediate need for a publication of the Mycenaean inscriptions in transcription. Until some time, the corpus of Mycenaean documents accompanied by the transliterated texts was available in printed form (Chadwick, et al., 1987). From then on, any work that has been done on the Mycenaean corpus is exhausted in the digitization of all published Mycenaean texts (Aurora, et al., 2013). The aim of digitization and online availability of the Mycenaean corpus is to make it reachable, searchable and constantly updated. In some cases, annotations and tools are provided that facilitate the study, analysis and understanding of the texts.

The previous work on the Mycenaean Linear B, either by semi-automatic methods (Packard, 1974; Duhoux, Palaima, & Bennet, 1989; Owens, 1997) or by using fully automated models (Luo, Cao, & Barzilay, 2019), exploits a database of individual Mycenaean words, i.e., the Mycenaean lexicon. In 1974, David Packard (1974) made a statistical analysis regarding the syllables of the Minoan Linear A and Mycenaean Linear B script based on the frequency of their appearance in initial, medial and final positions within the words of each language. For this purpose, Packard compiled all the individual words coming from the Mycenaean inscriptions available to him until then (“Mycenaeae Graecitatis Lexicon” (Morpugo, 1963)). In 1989 and 1997, Yves Duhoux (1989) and Gareth Owens (1997) respectively, repeated David Packard’s statistical analysis. In each study the dataset of words was updated with newly discovered inscriptions (Duhoux’s study for Linear B was based on “Linéaire B et Ordinateur Electronique” (Olivier, 1965) and Owens’ on “The Knossos tablets: A transliteration 5th edition” (Killen & Olivier, 1989)).

Recent works (Luo, Cao, & Barzilay, 2019) made use of more updated Mycenaean lexicons. Since all the Mycenaean inscriptions are available either in printed or electronic form, anyone can create their own dataset of Mycenaean words according to their own needs.

However, the use of isolated words is only useful at a level of linguistic morphological procedure, namely exporting morphological patterns. The methodical isolation of Mycenaean words, for the purpose of their study, has been carried out to a large extent both before (Kober, 1948; Ventris, 1988) and after (Ventris & Chadwick, 1953; Chadwick, 1967) the decipherment.

We contribute by presenting a dataset of organized groups of words derived from the Mycenaean texts, following the syntactic rules of the language. In a broader sense, we deal with all the Mycenaean language, the grammar of the language, and not just with the Mycenaean vocabulary.

The difficulties in studying the Mycenaean Linear B tablets, are obvious: a) The material state of the tablets, since most of them are broken, worn or burned, to a lesser or greater extent. b) The sententious nature of the Mycenaean inscriptions. Many of the tablets are made up of simple lists, revealing less than expected about their syntactic structure. c) The subject of the Mycenaean texts, which affected more the decipherment of the Linear B inscriptions. As they deal extensively with people and places, inevitably a large number of words contained in the tablets are proper names (Hooker, 1980).

Despite the aforementioned difficulties, the benefits are

(Packard, 1974), named-entity recognition (McCallum & Li, 2003), shallow parsing (Sha & Pereira, 2003; Sutton, Rohanimanesh, & McCallum, 2007), word segmentation (Peng, Feng, & McCallum, 2004), morphological segmentation (Kudo, Yamamoto, & Matsumoto, 2004), machine translation (Ravi & Knight, 2011b; Nuhn, Mauser, & Ney, 2012; Dou & Knight, 2012)).

- Contribute to decipherment of unknown or known scripts attributing unknown languages, like the Minoan script of the Second Millennium and even perhaps research concerning the “Neolithic script(s)” of the 6th Millennium, using computational techniques (e.g., HMM model based on Expectation-Maximization algorithm (Knight & Yamada, 1999; Knight, et al., 2006), Integer Programming (Ravi & Knight, 2008), Bayesian Inference (Snyder, Barzilay, & Knight, 2010; Ravi & Knight, 2011a), Neural Networks (Luo, Cao, & Barzilay, 2019)).

- Analyze and study linguistically the Mycenaean language (e.g. phonological, morphological, syntactic, semantic and factual analysis (Ruiperez & Melena, 1996; Hooker, 1980; Chadwick, 1976)).

3. Methodology

The creation of a valid and fully updated dataset is a difficult and laborious task. In this paper, our purpose is to design and raise awareness of a Mycenaean dataset of sequences of words in accordance with the principles of Mycenaean language, in order to be used for statistical

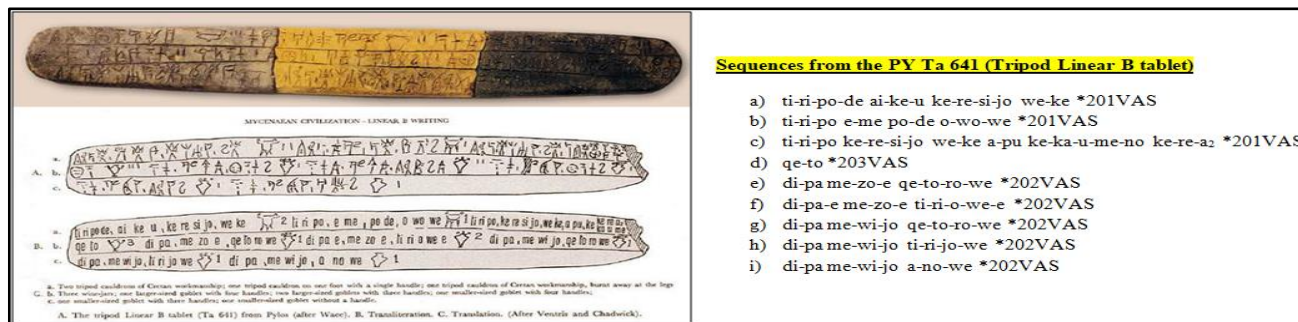


Figure 1: Left, the Tripod tablet from Pylos (PY Ta 641) (Ruiperez & Melena, 1996; Hooker, 1980). Right, the resulted sequences from the tripod tablet.

important. Such a dataset is not restricted to the examination of the language at a morphological level, but also gives valuable information on how this level interacts with other levels of analysis of Mycenaean language (syntax, semantics, etc.). In this way, we attempt to extract the conceptual content of the Mycenaean inscriptions. The rendering of the meaning of the Mycenaean texts enables predictions based on the Mycenaean language, such as the very important task of the complete restoration of the points which are illegible due to breakage or damage over the years.

A dataset consisting of sequences of Mycenaean words is intended for researchers who seek to:

- Extract inference focusing on statistical models which make probabilistic decisions (e.g. syllable correlations

research and study. Below it will be described in detail the creation of the Mycenaean dataset, what conventions were made and what rules were followed.

The digital dataset contains sequences of Mycenaean words, groups of from two to six and more syllables, in transcription, i.e., sequences of phonetic values as they have been identified and attributed to the syllables of the Mycenaean Linear B script. In order to generate this dataset all Linear B texts were used, mostly written in clay tablets (but also in clay vases, stamps and tags), found throughout the Mycenaean Greece (Knossos, Pylos, Mycenae, Thebes, Tiryns, etc.).

Each Mycenaean tablet is determined by two prefixes. The first consists of two capital letters, which indicate the

According to experts the inclusion of the Mycenaean ideograms in the dataset is necessary, since they played an

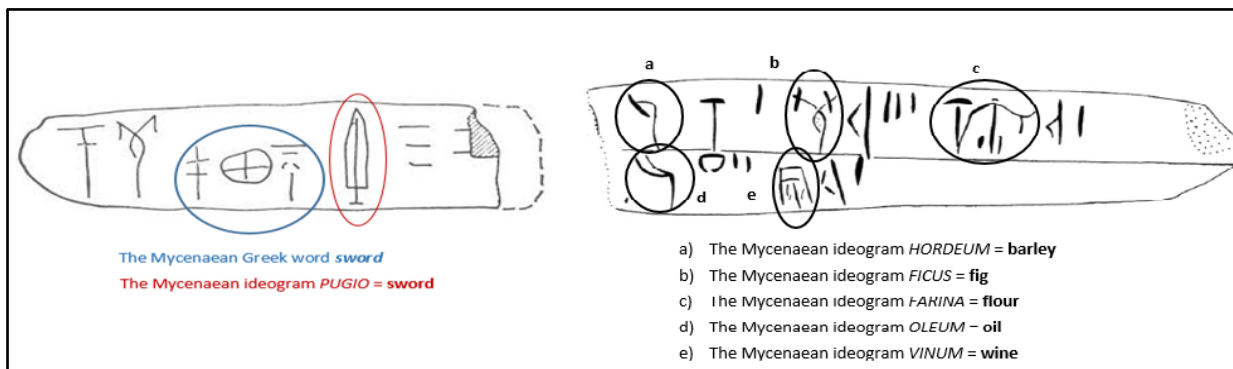


Figure 2: Left, in the Mycenaean tablet KN Ra 1540 (Hooker, 1980) the ideogram accompanies the word that describe. Right, the information of the Mycenaean tablet KN Fs 24 (Chadwick, et al., 1987) originates only from the ideograms.

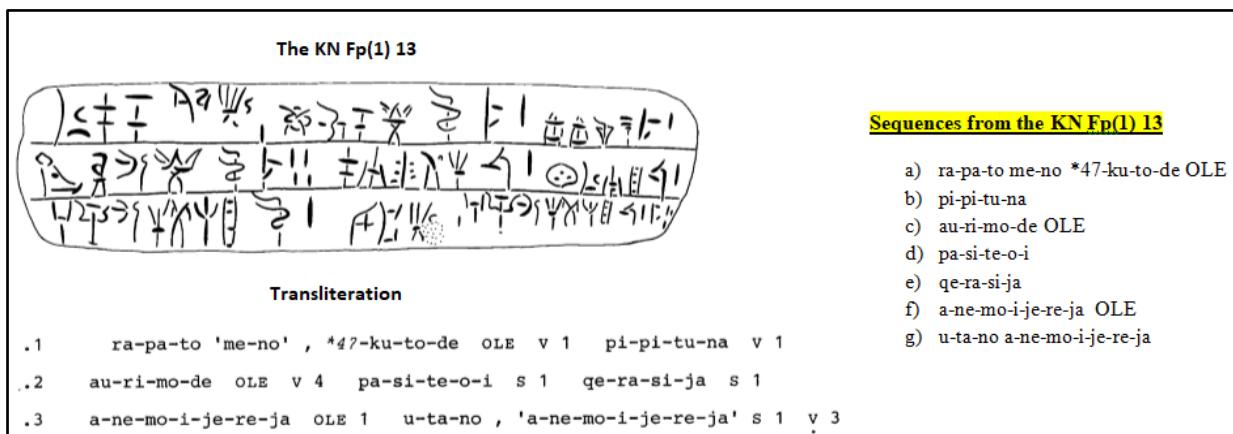


Figure 3: The Mycenaean tablet KN Fp 13 (Chadwick, et al., 1987) and the sequences extracted from it according to the methodology. The use of the divider for the separation of the words and the use of the numeric and metric signs for the separation of the sequences.

place of origin of the document (e.g., KN = Knossos, MY = Mycenae, PY = Pylos, etc. (Chadwick, 1976)). The second is a capital letter followed by a lowercase letter and serves to distinguish series of tablets, from which the expert understands the subject of the tablets (e.g., lists of persons, men or women, are marked with first letter A-, animal husbandry issues with C-, types of textiles with L-, etc. (Chadwick, 1976)). Finally, each tablet has its own ranking number (Chadwick, 1976). In this way, the tablet KN Fh 350 for example, is a document from Knossos, of the series Fh (oil records) and with ranking number 350 (Ruiperez & Melena, 1996).

The separation of the Mycenaean sequences was based on the following logic: The use of the divider (short upright line on the tablets, conventionally represented by a comma on the transliterated texts) as a word separation symbol and our hypothetical assumption that numeric and metric signs are followed by new sequences. This way, we sorted the signs which occur in groups, i.e., the words, as well as the ideograms. The numeric signs and the measurement units of weight and capacity were excluded from the dataset. The tablet of Figure 3 illustrates a representative example.

important role in the decipherment of the Linear B script. In fact, the tablet of Figure 1 confirmed Ventris' decipherment, since the design of the ideograms for the vessels and containers coincides with their description in the text (Ruiperez & Melena, 1996). The ideograms have semantic value; they represent a word and more specifically the meaning of a word. In some cases the ideograms are presented after the word they describe, confirming and identifying the information provided by the tablet. For example, in Figure 2 (left) the ideogram appears to simply repeat the message which is already satisfactorily expressed by the vocal part (Hooker, 1980). In other cases the information is derived exclusively from the ideograms, since the words are missing and these are the only elements of the tablets, see Figure 2 (right).

It is worth mentioning that the tablets of Figure 1 and 3 are a good case, not only because they offer several sentences but mainly because they are well-written and very well preserved. Unfortunately, most tablets are timeworn, broken or scrawled.

Considering the absence of strict syntactic rules in the Mycenaean language, it is difficult to create conventions of universal validity for the way of extracting sequences

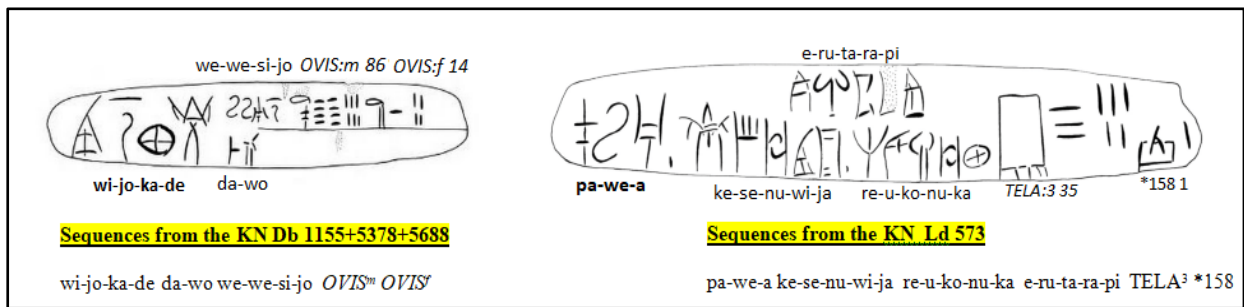


Figure 4: Left, the Mycenaean tablet KN Db 1155+5378+5688 (Chadwick, et al., 1987). Right, the Mycenaean tablet KN Ld 573 (Ruiperez & Melena, 1996; Chadwick, et al., 1987). Both attribute their content in a single sequence.

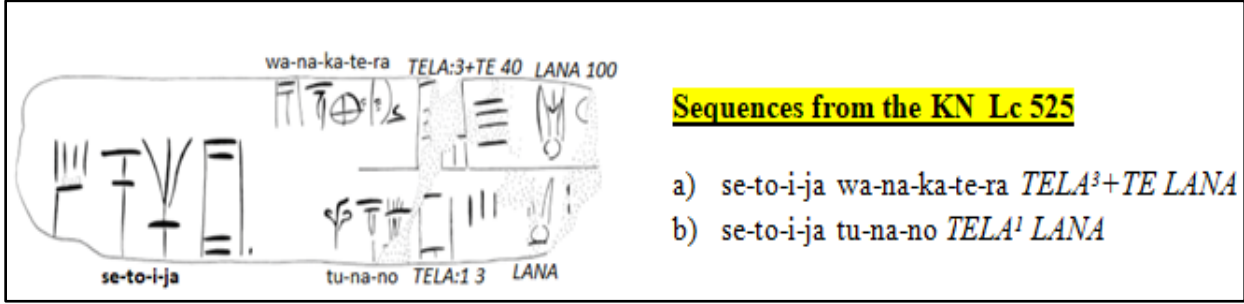


Figure 5: The Mycenaean tablet KN Lc 525 (Ruiperez & Melena, 1996; Chadwick, et al., 1987; Hooker, 1980). The content of the tablet is attributed in two sequences.

]to-so	No epigraphic evidence for the completeness of the word.
]to-so	Epigraphic evidence for a sign immediately preceding to.
]to-so	The word appears to be complete.

Figure 6: Conventions for illegible points of Mycenaean tablets (Chadwick, et al., 1987).

from all the tablets, making each tablet a special case. However, we attempt to systematize, as far as possible, the way of managing the less revealing tablets and the tablets in poor state.

A first category of tablets that presents particularity in the process of creating Mycenaean sequences is the one that follows the following structure: a group of large-sized syllables (word) on the left side of the tablet and two rows of groups of smaller syllables (words), usually separated by a dividing line, in the right side (Hooker, 1980), as shown on the tablets of figures 4 and 5. The handling of these tablets depends mainly on the series they belong, that is, on the content of the tablets. According to (Ruiperez & Melena, 1996), on the tablets with the above structure coming from series D- of Knossos, the text starts from the second line and continues on the first, as shown in Figure 4 (left). These tablets in their fullest form follow the pattern: [anthroponym (shepherd's name)] [toponym] [anthroponym (flock owner)]. The same happens with the tablets of the Ld series from Knossos, see Figure 4 (right). In those tablets, the words written in smaller characters are all neutral adjectives in plural which determine the word in big characters on the left (Hooker, 1980).

On the contrary, the tablets of the Lc series, indicate different content and are read differently despite they belong to the same category with those of Ld series, i.e., textile tablets. The words written in big characters in the

left part of the tablets are the locations where the textiles and wool were produced (Ruiperez & Melena, 1996; Hooker, 1980). Thus, from those tablets we usually gain two sequences (see Figure 5); the word in big characters refers to both rows of the right part of the tablets.

The proper way of creating the sequences, and how the Mycenaean tablets are read, is a key factor not only in capturing the meaning of the tablets but also in extracting the structure of the Mycenaean language. Those tasks are significantly impeded by syntactic inconsistencies, expected in lists that are not considered as sentences but have been structured in a fragmentary manner. So, because the Mycenaean texts are lists of personnel and goods (see for example a supermarket list) that were moved to the palaces, it is customary to differentiate among scribes, who present the same content and meaning by quoting words in a different order. Such variations should be taken into account and clarified in the dataset.

A second category that needs to be referenced is the one that includes tablets with illegible points, usually appearing on their broken edges, due to the damage and wear of the tablets or the misspelling and dysgraphia of the Mycenaean scribes. In the bibliography of the Mycenaean texts, there are general conventions for the indication of these points, as shown in Figure 6. These points could not be excluded from our dataset and below we will present the rules created for their inclusion.

The second case of Figure 6 is the most obvious. The symbols], -[and -[]- show that a syllable is definitely missing before or after or in-between a word. So, in this case we know, and it is also noticeable in the tablets, that there is certainly a syllable but this is not legible. In this case the syllable under examination is counted and is displayed with an * in our dataset, as shown in Figure 7. Although the difference in the first and third case of Figure 6 is not distinct, it is located in the gap between the

Melena, 1996; Hooker, 1980; Ventris & Chadwick, 1973) is the verb *have* in the third person singular in the simple present tense and the word “*ki-u-ro*” (Ventris & Chadwick, 1973) is an anthroponym (a person’s name). In some cases the number of missing symbols may be more than one. But because it is impossible to know the exact number of missing syllables, a good start is to declare that certainly one syllable is missing. One last thing that should be mentioned for this category of tablets


Mycenaean Tablet X 44	Transliterated Text	Sequences from the KN X 44
	.A wi-ri-[.B ku-ja-ro / qa-ra [ku-ja-ro qa-ra wi-ri-*
	Comment .A Traces at right consistent with za.	

Figure 7: In the Mycenaean tablet KN X 44 (Chadwick, et al., 1987) definitely missing at least one syllable on its right damaged part.



Mycenaean Tablet KN Dq(3) 45	Transliterated Text	Sequences from the KN Dq(3) 45
	.a] a-no-qo-ta-o [.b]mo / e-ra ovis ^m [*-mo e-ra a-no-qo-ta-o OVIS ^m
Mycenaean Tablet KN DI 47	Transliterated Text	Sequences from the KN DI 47
	.1]e-ke , e-u-da-i-ta ovis ^f 39[.2]ki-u-ro / su-ki-ri-ta-pi o ki ^t OVIS 15 [a) e-ke e-u-da-i-ta OVIS ^f b) ki-u-ro su-ki-ri-ta-pi o-pe-ro ki OVIS

Figure 8: Up, on the damaged left side of the Mycenaean tablet KN Dq 45 (Chadwick, et al., 1987) is missing at least one syllable. Down, the Mycenaean tablet KN DI 47 (Chadwick, et al., 1987) is considered complete, no syllable is missing from its damaged left part.

symbol and the syllable. In the third case the gap is larger than in the first. Thus, the symbols], [and [] in the first case indicate that a syllable may be missing before or after or in-between a word. The same symbols in the third case show that the word is probably complete. Once we discern the first from the third case in a tablet, the rest of the process lies in the decision we will take for the first case. The answer is given by the content of the tablet itself. As demonstrated in Figure 8, for the first tablet (up) was decided that a syllable is missing before the syllable “*mo*”, even if there is no epigraphic evidence. Such a decision seems perfectly reasonable since there is no Mycenaean word with a single syllable. Even in the case that the syllable “*mo*” functioned as an ideogram, it should be followed by numbers. On the contrary, for the second tablet (down), was decided that the words “*e-ke*” and “*ki-u-ro*” are complete, since they are words of the Mycenaean vocabulary. The word “*e-ke*” (Ruiperez &

are the notes appearing in some of them, regarding the aforementioned illegible points, as illustrated in Figure 7. In this tablet, the bibliography (Chadwick, et al., 1987) provides a note on the illegible syllable. These notes are also taken into account in our dataset. Each missing syllable appears in the dataset with an * and also through the creation of probabilistic vectors, capable of capturing such meanings (see section 5).

4. Dataset Description

We aim to create a dataset of Mycenaean sequences originating from all the available Mycenaean inscriptions found until today. Our sources for the Mycenaean inscriptions derive from their most recent bibliography as well as from books containing annotated anthology on Mycenaean texts (Ruiperez & Melena, 1996; Hooker, 1980; Chadwick, The Mycenaean World, 1976).

The file where all the Mycenaean sequences are compiled is composed by 19 sheets, each of which presents a category of Linear B inscriptions by giving their location, abbreviated with the first two letters of the toponym, accompanied by their series. The inscriptions on each sheet are arranged in increasing ranking numbers, which follow the taxonomic prefixes. The majority of the tablets give more than one sequence. For example, the longest Knossos tablet (Figure 10, right) provides 70 sequences, most of which are men's names (Ventris & Chadwick, 1973). Some tablets because of the great destruction they have suffered are not capable of providing us with sufficient information for their own study and processing

The total number of sequences amounts to more or less 10000 and come from about 6000 inscriptions. From these 6000 about one third are fully processed. The rest are expected to be completed in the following months. In Figure 9 we present some sequences from the sheet with name 'KN SERIES A- & B'. For example, the tablet Ak 627 from Knossos gives 5 sequences. This tablet records a workgroup on the Cretan locale *Da-*22-to*, which belongs to *Anorxo* (personal name in genitive case) (Ruiperez & Melena, 1996). The ideograms and the phonetic signs used as ideograms are printed in italic capitals. The adjuncts that modify the meaning of ideograms are in lower-case italic placed in the same cell

Ak(2) 627 + 7025 + fr.	da	*22	to		a	no	zo	jo		TA		DA		MUL		pediMUL
Ak(2) 627 + 7025 + fr.	da	*22	to		a	no	zo	jo		ko	wa		me	zo	e	
Ak(2) 627 + 7025 + fr.	da	*22	to		a	no	zo	jo		ko	wa		me	wi	jo	e
Ak(2) 627 + 7025 + fr.	da	*22	to		a	no	zo	jo		ko	wo		me	zo	e	
Ak(2) 627 + 7025 + fr.	da	*22	to		a	no	zo	jo		ko	wo		me	wi	jo	e
Ap 628 + 5935	*	i	ja		a	ke	wo			do	e	ra		MUL		
Ap 628 + 5935	*	ro		do	e	ra			MUL							
Ap 628 + 5935	*	ro		di	qa	ra	*									
Ap 628 + 5935	*	ne	o		do	e	ra			MUL						
Ap 629	tu	ni	ja		tuMUL			nediMUL		koMUL						
Ap 629	ri	jo	no		tuMUL											
Ap 629	ri	jo	no		ko	wo										
Ap 629	do	ti	ja		tuMUL			nediMUL								
Ak(2) <631>																
Ai 632	*	ta	ra2			MUL										

Figure 9: Sample of the Mycenaean dataset. Sequences derived from the tablets KN Ak 627, KN Ap 628, KN Ap 629, KN Ak 631 and KN Ai 632 (Chadwick, et al., 1987).

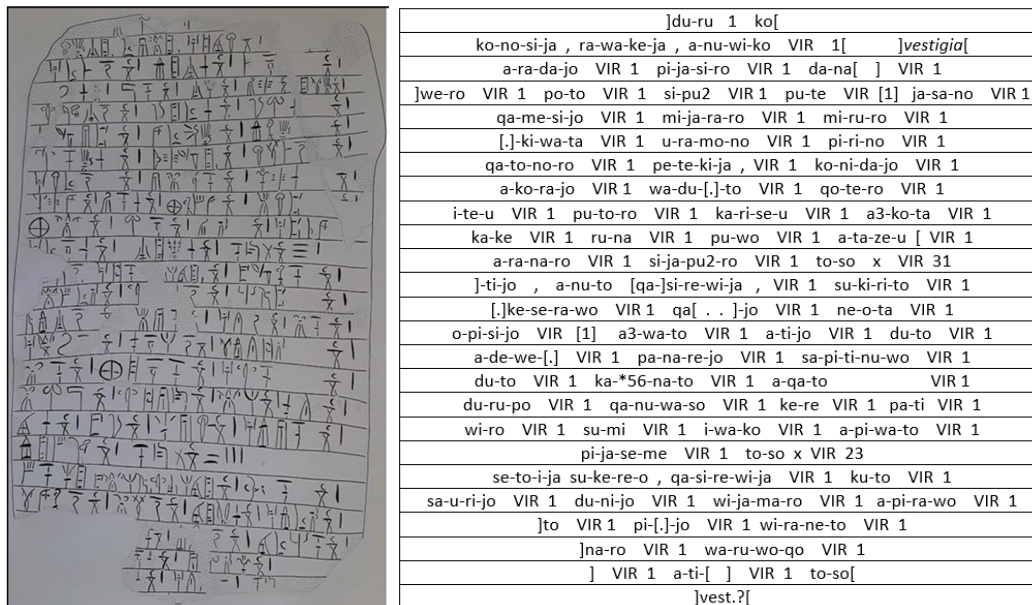


Figure 10: Left, the Mycenaean Tablet KN As 1516 (Chadwick, et al., 1987). Right, the transliterated text of the KN As 1516 (Chadwick, et al., 1987).

(huge number of probabilities regarding the content of missing pieces). These tablets appear in the dataset with empty cells. Each cell hosts a Mycenaean syllable, including the space, or a Mycenaean ideogram. The phonetic signs, usually abbreviations of Mycenaean words, that are used as adjuncts to modify the value of a following ideogram are placed in the same cell with the ideogram and are counted as separate ideograms. Thus, the Mycenaean sequences consist of around 90 syllables and over 100 ideograms.

and before the ideograms. The unidentified Mycenaean syllables appear with an *. Finally, the tablet Ak 631 does not provide any significant information.

5. Experiment

An initial experiment on the dataset concerns the illegible points of the series A and B of Knossos, which include tablets referring to lists of personnel. The aim is to predict the phonetic values of the unidentified signs of the Mycenaean tablets of categories A and B of Knossos by

training a CRF model. Figure 10 (left) presents the damaged parts of the tablet resulting in incomplete and ambiguous words as shown in Figure 10 (right). Here we make predictions for completing these missing parts. This experiment will be the initial step towards the clarification of the ambiguities that remain in the transliteration of the Mycenaean Linear B and ultimately towards the decipherment of the Minoan script (a Phonetic Value Attribution task).

MULIER (WOMAN), since these tablets are intended for personnel record. Therefore, in total were collected 98 different symbols which constitute the model's observations (O). Each syllable attributes its own phonetic value and each ideogram relates to its own type of object. So, the number of model's labels (L) is also 98. With this labeling, an example sentence derived from KN Ai 63, was presented as shown in Table 1. To define the relationships among the observations and

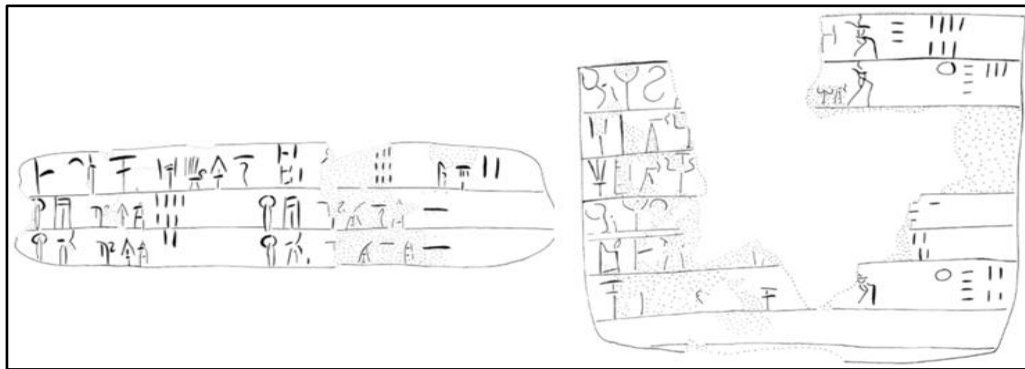


Figure 11: Left, the well preserved Mycenaean tablet KN Ak 627 (Chadwick, et al., 1987). Right, the heavily damaged Mycenaean tablet KN B 164 (Chadwick, et al., 1987).

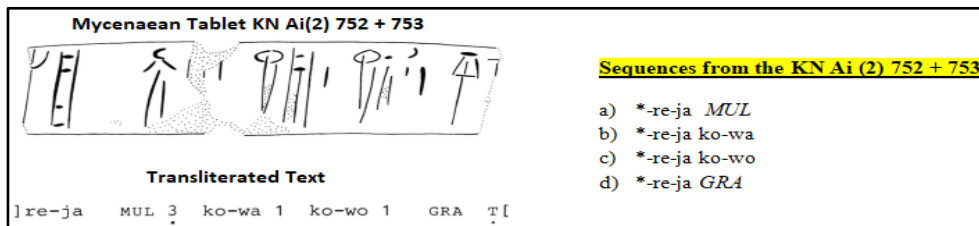


Figure 12: The damaged Mycenaean tablet KN Ai 752+753 (Chadwick, et al., 1987) and the resulted sequences.

In this experiment we used all the tablets of category A and B from Knossos, 308 in number, from which we obtained 651 sequences according to the methodology of section 3. From these tablets only about 20% are complete, Figure 11 (left). The rest are from slightly to fairly damaged, Figure 11 (right). The category we are studying refers to the annual staffing, that is, it records the number of persons, men (ideogram *VIR*), women (ideogram *MUL*), girls (ko-wa) and boys (ko-wo), employed in various activities.

labels, we need to choose the set F of feature functions $f_k(y_t, y_{t-1}, x_t)$ (Sutton & McCallum, 2012). In our case of linear-chain CRF there are two essential kinds of features. The first kind examines the pairs of adjacent labels $f_{ij}^{LL}(y_t, y_{t-1}) = \mathbf{1}_{\{y_t=i\}}\mathbf{1}_{\{y_{t-1}=j\}} \forall i, j \in \mathcal{Y}$, where \mathcal{Y} are the different labels of the model, and it is called the *label-label features*. For this problem, there are 98 different labels, so there are 9,604 label-label features. The second kind examines each label-observation pair $f_{io}^{LO}(y_t, x_t) = \mathbf{1}_{\{y_t=i\}}\mathbf{1}_{\{x_t=o\}} \forall i \in \mathcal{Y}, o \in \mathcal{O}$, where \mathcal{O} is the set of all unique syllables and ideograms that appear in this dataset, and it is called the *label-observation features*. In our dataset there are 98 different symbols (syllables and ideograms), so there are 9,604 label-observation features. As for the points of the Mycenaean tablets under evaluation, which appear with an * in our dataset, since we don't know their labels, but we look for them, they appear with vectors of dimension 98. Each element of the vector expresses the probability of each label being present. For example, in the tablet of Figure 12, which appears broken in its left and right part, there is certainly one or more syllables before the syllable 're'. Consequently in the vector describing the unknown symbol, the pdf will be uniform in the phonetic values attributed by the 77 syllables, since this is impossible to be an ideogram. In this case the label-label features will have the form $f_{ij}^{LL}(y_t, y_{t-1}) = \mathbf{1}_{\{y_t=re\}}\mathbf{p}(j)_{\{y_{t-1}=j\}} \forall j \in$

t	y_t (labels)	x_t (observations)
0	pe	𐀀
1	se	𐀁
2	ro	𐀂
3	jo	𐀃
4	blank	
5	e	𐀄
6	e	𐀄
7	si	𐀅
8	blank	
9	MUL	𐀆

Table 1: Sample sequence with labels and observations.

To achieve our goal we trained a linear-chain CRF model in Matlab. In the 'KN SERIES A- & B' dataset, the different syllables are 78 (including the blank) and the different ideograms of this category are 20, the most representative of which are the *VIR* (MAN) and the

phonetic value, where $p(j)=1/77=0.013$ and $f_{ij}^{LL}(y_t, y_{t-1}) = \mathbf{1}_{\{y_t=re\}} \mathbf{0}_{\{y_{t-1}=j\}} \forall j \in \text{object}$.

Before conducting the experiment, we need to examine the reliability of our model. So, to validate our model we firstly defined a test set of tablets without damaged points. Since the Mycenaean Linear B script has been deciphered, the related label sequences are known and used as ground-truth. Thus, we trained the model in 622 sequences and tested in 3 different tablets, Ak 627+7025+fr. (tablet formed by the fragment association), Ai 739 and As 1517, each offering 5, 3 and 21 sequences respectively.

The percentage of incorrect predictions was estimated to be 0.8417 for the tablet KN As 1517, 0.2174 for the tablet KN Ai 739 and 0.0513 for the tablet KN Ak 627+7025+fr.

The ground truth labels of the Mycenaean tablet Ak 627+7025+fr.	
da-*22-to a-no-zo-jo	TA DA MUL <i>pediMUL</i>
da-*22-to a-no-zo-jo	ko-wa me-zo-e
da-*22-to a-no-zo-jo	ko-wa me-wi-jo-e
da-*22-to a-no-zo-jo	ko-wo me-zo-e
da-*22-to a-no-zo-jo	ko-wo me-wi-jo-e
The estimated labels of the Mycenaean tablet Ak 627+7025+fr.	
da-*22-to a-no-zo-jo	<i>VIR VIR MUL VIR</i>
da-*22-to a-no-zo-jo	ko-wa me-zo-e
da-*22-to a-no-zo-jo	ko-wa me-wi-jo-e
da-*22-to a-no-zo-jo	ko-wo me-zo-e
da-*22-to a-no-zo-jo	ko-wo me-wi-jo-e
The ground truth labels of the Mycenaean tablet Ai 739	
ra-su-to a-ke-ti-ri-ja	<i>MUL</i>
ra-su-to ko-wa	
ra-su-to ko-wo	
The estimated labels of the Mycenaean tablet Ai 739	
ra-su-to	<i>ko-we-ko-we-ja MUL</i>
ra-su-to	ko-wo
ra-su-to	ko-wo

Table 2: Testing the trained model into the Mycenaean tablets Ak 627+7025+fr. and Ai 739

Table 2 shows the output compared to the ground-truth for two tablets. The KN As 1517 tablet, to which belongs the highest error, records the male workforce at Knossos, largely recorded by their names. Simply listing a multitude of individual male names and the absence of their repetition complicates the corpus, greatly hampering our predictions. Most of these names occur in the Mycenaean documents only once. Therefore, the respective f_{ij}^{LL} functions had low weight, which explains why we got such a high error rate.

On the contrary, the women and children at Knossos (Ai and Ak series) were identified either by their occupation (trade-name), which sometimes came with the location (place-name) where they worked, or by a toponymic adjective (ethnic-name). In some cases there was added or substituted a man's name (personal-name), which might be in the nominative or genitive, and referred to the owner of a group of workers (Ventris & Chadwick, 1973). In this case, despite their conciseness the tablets provide linguistic information (e.g., the genitive of male names or the feminine of the ethnic adjective derive from a place-name) and cultural information (e.g., the toponyms, the professions, the institutions and the societal hierarchy).

The experiment was carried out in a category of tablets, which is rather small in size (651 sequences derived from 308 tablets). Typically, clay tablets were annual draft censuses used for economic-administrative notes and were destroyed every financial year to reuse clay for new recordings. Only a small amount of such tablets survived. For example, since our experimental sample is an annual aggregate record of Knossos' personnel, the word 'a-ke-ti-

ri-ja' (female workers in textile processing) in tablet Ai 739 appears at most one or two times. This had resulted in the model not being able to correctly predict this occupational term (Table 2), since the respective f_{ij}^{LL} functions of this word, had low weight.

On the other hand there are words (e.g., ko-wo, ko-wa, me-zo, me-wi-jo and me-u-jo) and ideograms (e.g., *VIR* and *MUL*) that are frequent in the dataset. The respective feature functions are assigned high weights. On the contrary, as seen in Ak 627 tablet (Table 2), the model fails to predict ideograms *DA*, *TA* and *pediMUL* since they rarely appear in the sample and classify them as *VIR* which is repeated so many times in the dataset.

Experiment Our primary purpose is to predict the labels attributed by the missing symbols of the illegible parts of the Mycenaean tablets. We applied the model in the sequences derived from the damaged Mycenaean tablets. Table 3 shows the results of the CRF model only on the damaged parts of the Mycenaean tablet As 1516. These results coincide with the estimation of the experts on the missing symbols (see (Chadwick et al, 1987), Volume 2, page 149). For example, for the missing syllable of the word [-]ki-wa-ta, the book suggests: "a-ki-wa-ta or a3-ki-wa-ta possible". The output of our model agrees with the first estimate. The same goes for the word wa-du-[-]to. The experts suggests: "wa-du-ni-to or (less likely) wa-du-sa-to". The CRF agrees with the most likely estimate.

ko-no-si-ja , ra-wa-ke-ja , a-nu-wi-ko	<i>VIR</i> []	<i>vestigia</i> [
a-ra-da-jo	<i>VIR</i> 1 pi-ja-si-ro	<i>VIR</i> 1 da-na-ro	<i>VIR</i> 1		
pe-we-ro	<i>VIR</i> 1 po-to	<i>VIR</i> 1 si-pu2	<i>VIR</i> 1 pu-te	<i>VIR</i> [1] ja-sa-no	<i>VIR</i> 1
qa-me-si-jo	<i>VIR</i> 1 mi-ja-ra-ro	<i>VIR</i> 1 mi-ru-ro	<i>VIR</i> 1		
a-ki-wa-ta	<i>VIR</i> 1 u-ra-mo-no	<i>VIR</i> 1 pi-ri-no	<i>VIR</i> 1		
qa-to-no-ro	<i>VIR</i> 1 pe-te-ki-ja ,	<i>VIR</i> 1 ko-ni-da-jo	<i>VIR</i> 1		
a-ko-ra-jo	<i>VIR</i> 1 wa-du-ni-to	<i>VIR</i> 1 qa-te-ro	<i>VIR</i> 1		
i-te-u	<i>VIR</i> 1 pu-to-ro	<i>VIR</i> 1 ka-ri-se-u	<i>VIR</i> 1 a3-ko-ta	<i>VIR</i> 1	
ka-ke	<i>VIR</i> 1 ru-na	<i>VIR</i> 1 pu-wo	<i>VIR</i> 1 a-ta-ze-u	[<i>VIR</i> 1
a-ra-na-ro	<i>VIR</i> 1 si-ja-pu2-ro	<i>VIR</i> 1 to-so	x	<i>VIR</i> 31	
ra-ti-jo ,	a-nu-to	qa-si-re-wi-ja ,	<i>VIR</i> 1 su-ki-ri-to	<i>VIR</i> 1	
a-ke-se-ra-wo	<i>VIR</i> 1 qa .	wi-jo	<i>VIR</i> 1 ne-o-ta	<i>VIR</i> 1	
o-pi-si-jo	<i>VIR</i> [1] a3-wa-to	<i>VIR</i> 1 a-ti-jo	<i>VIR</i> 1 du-to	<i>VIR</i> 1	
a-de-we-ro	<i>VIR</i> 1 pa-na-re-jo	<i>VIR</i> 1 sa-pi-ti-nu-wo	<i>VIR</i> 1		
du-to	<i>VIR</i> 1 ka-*56-na-to	<i>VIR</i> 1 a-qa-to	<i>VIR</i> 1		
du-ru-po	<i>VIR</i> 1 qa-nu-wa-so	<i>VIR</i> 1 ke-re	<i>VIR</i> 1 pa-ti	<i>VIR</i> 1	
wi-ro	<i>VIR</i> 1 su-mi	<i>VIR</i> 1 i-wa-ko	<i>VIR</i> 1 a-pi-wa-to	<i>VIR</i> 1	
pi-ja-se-me	<i>VIR</i> 1 to-so	x	<i>VIR</i> 23		
se-to-ija	su-ke-re-o ,	qa-si-re-wi-ja	<i>VIR</i> 1 ku-to	<i>VIR</i> 1	
sa-u-ri-jo	<i>VIR</i> 1 du-ni-jo	<i>VIR</i> 1 wi-ja-ma-ro	<i>VIR</i> 1 a-pi-ra-wo	<i>VIR</i> 1	
*18-to	<i>VIR</i> 1 pi-ri-jo	<i>VIR</i> 1 wi-ra-ne-to	<i>VIR</i> 1		
ru-na-ro	<i>VIR</i> 1 wa-ru-wo-qa	<i>VIR</i> 1			
] <i>VIR</i> 1 a-ti-ja	<i>VIR</i> 1 to-so[

Table 3: Missing symbols estimation in the Mycenaean tablet KN As 1516.

6. Conclusions

We presented a dataset of Mycenaean Linear B sequences using the corpus of tablets discovered so far. We run initial experiments using a CRF to verify that we find the phonetic correspondence to symbols and to predict missing symbols. The application of the model, in conjunction with its evaluation by experts in the future, is a focused effort to complete the damaged Mycenaean tablets and aspires to contribute in the decipherment efforts of ancient scripts. The dataset will become fully available online once completed and reviewed by experts.

Acknowledgement Co-financed by the EU and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-00502 - MuseLearn).

References

- Aurora, F., Nesøen, A., Nedić, D., Løken, H., & Bersi, A. (2013). *DAMOS - Database of Mycenaean at Oslo University of Oslo*. Retrieved 2019, from <https://www2.hf.uio.no/damos/>
- Brent, D. (2010). Introduction to the Aegean pre-alphabetic scripts. *Kubaba I*, pp. 38-61.
- Chadwick, J. (1960). *The Decipherment of Linear B* (2nd ed.). Cambridge University Press.
- Chadwick, J. (1967). *The decipherment of Linear B*. Cambridge University Press.
- Chadwick, J. (1976). *The Mycenaean World*. Cambridge/London/ NewYork/ Melbourne: Cambridge University Press.
- Chadwick, J. (1987). *Linear B and Related Scripts*. University of California Press.
- Chadwick, J., Godart, L., Killen, J. T., Olivier, J. P., Sacconi, A., & Sakellarakis, I. A. (1987). *Corpus of Mycenaean Inscriptions from Knossos: Volumes 1-4*. Cambridge University Press.
- Chadwick, J., Godart, L., Killen, J. T., Olivier, J. P., Sacconi, A., & Sakellarakis, I. A. (n.d.). *Corpus of Mycenaean Inscriptions from Knossos: Volumes 1-4*. Cambridge University Press.
- Corlett, E., & Penn, G. (2010). An exact A* method for deciphering letter-substitution ciphers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (pp. 1040-1047).
- Dou, Q., & Knight, K. (2012). Large Scale Decipherment for Out-of-Domain Machine Translation. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.
- Duhoux, Y., & Morpurgo Davies, A. (2008). *A Companion to Linear B: Mycenaean Greek Texts and their World* (Vol. 1). Peeters Louvain-La-Neuve.
- Duhoux, Y., Palaima, T. G., & Bennet, J. (1989). Le linéaire A: problèmes de déchiffrement. In *Problems in Decipherment* (pp. 59-119). Peeters, Louvain-la-Neuve.
- Hooker, J. T. (1980). *Linear B: An introduction*. Bristol Classical Press.
- Killen, J. T., & Olivier, J.-P. (1989). *The Knossos tablets: A transliteration (Minos Supplement 11)* (5th ed.). Salamanca, España: Ediciones Universidad de Salamanca.
- Knight, K., & Yamada, K. (1999). A computational approach to deciphering unknown scripts. In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Unsupervised Learning in Natural Language Processing*, (pp. 37-44).
- Knight, K., Magyesi, B., & Schaefer, C. (2011). The Copiale cipher*. In *the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, (pp. 2-9).
- Knight, K., Nair, A., Rathod, N., & Yamada, K. (2006). Unsupervised analysis for decipherment problems. In *Proceedings of the International Conference on Computational Linguistics (COLING)/Association for Computational Linguistics (ACL) 2006 Main Conference Poster Sessions*, (pp. 499-506).
- Kober, A. E. (Jan-Mar 1948). The Minoan Scripts: Fact and Theory. *American Journal of Archaeology*, 52 (No 1), 82-103.
- Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Luo, J., Cao, Y., & Barzilay, R. (2019). Neural Decipherment via Minimum-Cost Flow: from Ugaritic to Linear B. *Association for Computational Linguistics*. Florence.
- McCallum, A., & Li, W. (2003). Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL)*.
- Morpugo, A. (1963). *Mycenaeae Graecitatis Lexicon*. Rome: In Aedibus Athenaein.
- Nuhn, M., & Knight, K. (2014). Cipher type detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1769-1773). Doha, Qatar: Association for Computational Linguistics.
- Nuhn, M., & Ney, H. (2013). Decipherment complexity in 1:1 substitution ciphers. In *the 51st Annual Meeting of the Association for Computational Linguistics*, (pp. 615-621).
- Nuhn, M., Mauser, A., & Ney, H. (2012). Deciphering foreign language by combining language models and context vectors. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 156-164).
- Olivier, J.-P. (1965). Linéaire B et ordinateur électronique. In *L'Antiquité Classique 34* (pp. 387-397). Brussels.
- Owens, G. A. (1997). "Computer Techniques in the Study of Minoan Linear A script" (1971-1996) Back to the Future? In *Kritika Daidalika: Evidence for the Minoan Language: Selected Essays in Memory of James Hooker on the Archaeology, Epigraphy and Philology of Minoan and Mycenaean Crete* (pp. 175-186). Amsterdam: Adolf M. Hakkert.
- Owens, G. A. (2007). *Labyrinth: Scripts and Languages of Minoan and Mycenaean Crete*. Heraklion: Centre for Cretan Literature.
- Owens, G. A. (2008-2018). *Tei of Crete - Daidalika*. (PASIPHAЕ Research and Development of

- Telecommunications Systems Laboratory, Department of Informatics Engineering, at the "DAIDALIC" Technological Educational Institute of Crete, Hellas.) Retrieved from <https://www.teicrete.gr/daidalika/>
- Packard, D. W. (1974). *Minoan Linear A*. University of California Press.
- Papavasileiou, K. (2010). *BACHELOR's DISSERTATION: Statistical data analysis of Linear A compared with the Linear B using Pattern Recognition*. Submitted to the department of Informatics Engineering of School of Applied Technology of Technological Educational Institute of Crete.
- Papavasileiou, K. (2014). *MASTER's THESIS: Pattern Recognition and Computational Intelligence in Scripts and Languages of Minoan and Mycenaean Crete From Mycenaean Linear B to the Minoan Phaistos Disk?*. Submitted to the department of Informatics Engineering of School of Applied Technology of Technological Educational Institute of Crete.
- Papavasileiou, K., Papadourakis, G., & Owens, G. (2011). Statistical Analysis of Minoan Linear A in Relation to Mycenaean Linear B using Pattern Recognition. *7th International Scientific Conference: New Horizons in Industry, Business and Education, (NHIBE 2011)*. Chios.
- Papavasileiou, K., Papadourakis, G., & Owens, G. (2013). Artificial Intelligence in Scripts and Languages of Minoan and Mycenaean Crete. From Mycenaean Linear B to the Minoan Phaistos Disk? *8th International Scientific Conference: New Horizons in Industry, Business and Education, (NHIBE 2013)*. Chania.
- Peng, F., Feng, F., & McCallum, A. (2004). Chinese Segmentation and New Word Detection using Conditional Random Fields. *Proceedings of the International Conference on Computational Linguistics (COLING)*.
- Ravi, S., & Knight, K. (2008). Attacking decipherment problems optimally with low-order n-gram models. *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 812-819).
- Ravi, S., & Knight, K. (2009). Learning phoneme mappings for transliteration without parallel data. *In Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 37-45).
- Ravi, S., & Knight, K. (2011a). Bayesian Inference for Zodiac and Other Homophonic Ciphers. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (pp. 239-247).
- Ravi, S., & Knight, K. (2011b). Deciphering foreign language. *In the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (pp. 12-21).
- Robinson, A. (2002). *The Man Who Deciphered Linear B: The Story of Michael Ventris*. Thames & Hudson.
- Robinson, A. W. (2002). *Lost Languages: The Enigma of the World's Undeciphered Scripts*.
- Ruiperez, M. S., & Melena, J. L. (1996). *The Mycenaean Greeks*. Athens: Kardamitsa.
- Sha, F., & Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. *Proceedings of the Conference on Human Language Technology and North American Association for Computational Linguistics (HLT-NAACL)*.
- Snyder, B., Barzilay, R., & Knight, K. (2010). A statistical model for lost language decipherment. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (pp. 1048-1057).
- Sutton, C., & McCallum, A. (2012). *An Introduction to Conditional Random Fields*. Now Publishers.
- Sutton, C., Rohanimanesh, K., & McCallum, A. (2007). Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. *Machine Learning Research*, 8, 693-723.
- Ventris, M. (1988). 1951-1952 " Work notes on Minoan language research" nos.1-20. In A. Sacconi (Ed.), *Work notes on Minoan language research and other unedited papers (Volume 90 of Incunabula Graeca)* (pp. 135-333). Rome: Edizioni dell'Ateneo.
- Ventris, M., & Chadwick, J. (1953). Evidence for Greek dialect in the Mycenaean archives. *Journal of Hellenic Studies (JHS)*, 73, 84-103.
- Ventris, M., & Chadwick, J. (1973). *Documents in Mycenaean Greek*. Cambridge [England] University Press.