A Vowel Harmony Testing Algorithm to Aid in Ancient Script Decipherment

Peter Z. Revesz
Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, NE, USA
revesz@cse.unl.edu
0000-0002-1145-1283

***Abstract*— Previous algorithms for deciphering lost languages assumed knowledge of cognate languages. Since that is not always possible a priori, this paper develops an algorithm that can test an important feature of the underlying language, namely the presence of vowel harmony in root words. The new algorithm detects this type of vowel harmony in the Minoan language, thereby greatly narrowing its possible set of cognate languages.**

Keywords— decipherment, script, syllable graph, vowel harmony

I. Introduction

There is still no algorithmic method to decipher the scripts of truly unknown languages. Snyder et al. [22] proposed a general method of decipherment of lost languages, but it assumes knowledge that the lost language is a cognate of some known language and their respective scripts are also cognates. Snyder et al. [22] gave as an example the decipherment of Ugaritic under the prior knowledge that it is a cognate of Hebrew.

The problem is that for many other scripts, such as the Cretan Hieroglyphs [15] and Linear A [8] that express the unknown Minoan language, there is no clear cognate candidate. The island of Crete is readily accessible by sea from three continents virtually from any direction. No wonder that even a single inscription, the Phaistos Disk, was claimed to be translated into dozens of languages, including Basque (Gordon [9]), Georgian (Kvashilava [11]), Greek (Faucounau [5] and Fischer [6]), Hittite (Georgiev [7]), Luwian (Achtenberg et al. [2]), Semitic (Aartun [1], Gordon [10]), Sumerian (Massey and Massey [13]), and Proto-Ugric (Revesz [17]), to give just a small sample of the numerous translation attempts.

The great uncertainty regarding which languages may be cognates of the Minoan language prevents applying the method of Snyder et al. [22], which requires some known cognates. Therefore, from the computational linguistics point of view we need a preprocessing step that narrows down the set of possible cognate languages by the use of new algorithmic methods. We address that question in this paper.

By testing the undeciphered script on certain language features, the set of possibilities can be narrowed down. For example, if the unknown language is agglutinative, then its cognates will be agglutinative too. Duhoux [4] already showed that the Minoan language is agglutinative. Unfortunately, that is not enough because Basque, Georgian, Sumerian and Finno-Ugric are all agglutinative languages. We need to search for some other computationally testable feature that distinguishes these languages too. After searching, we found that *vowel harmony* is a feature that can be tested algorithmically.

The rest of this paper is organized as follows. Section 2 reviews the concept of vowel harmony and describes an algorithm that checks for vowel harmony. Section 3 considers Cretan Hieroglyphic inscriptions. Section 4 discusses the results. Finally, Section 5 gives some conclusions and describes future work.

II. Description of the Algorithm

Vowel harmony of some kind is a feature of many languages, including Korean, Mongolian, Turkic and Uralic languages [12]. A particular type of vowel harmony occurs when all the vowels of the root words are back vowels, as in 'coconut' or front vowels, as in 'kiwi.' We show below that we can test whether syllabic inscriptions of an unknown language indicate the presence of vowel harmony. We assume that the script is a syllabary where each symbol consists of a single vowel or a CV type syllable, where C is a consonant and V is a vowel. An example of such a syllabary is the Linear B script, which is the oldest form of Greek writing.

Most epigraphic experts assume that the Linear A and the Cretan Hieroglyphic scripts, are similar syllabaries. They also consider the Phaistos Disk to be a form of Cretan Hieroglyphs, albeit a unique artifact because it was printed by a set of seals representing the writing symbols. The inscription is written in a spiral on both sides of the clay disk with 31 well-separated blocks on side A and 30 blocks on side B. There is debate whether the disk is to be read from the center to the outside (or left-to-right) or from the outside to the center (or right-to-left).

Given a document library λ, a script σ, and an integer threshold value τ, our algorithm can be described as follows:

Algorithm Vowel-Harmony-Check(λ, σ, τ)

1. **Find frequencies of pairs and triplets of symbols**. If we have n different symbols in the script, then this step can be done by initializing the two frequency matrixes $A[1 \dots n][1 \dots n]$ and $B[1 \dots n][1 \dots n][1 \dots n]$ to zero, then while scanning the available documents we update the entry for A[i, j] (and B[i, j, k]) whenever the i^{th}, j^{th} (and k^{th}) symbols are found adjacently.
2. **Find hypothetical root words.** We consider only sequences which have frequencies ≥ τ. Out of the frequent pairs and triplets, we eliminate those that only occur at the end of root words.

3. **Create an adjacency graph**. We create a vertex for each symbol of σ. Then we connect each pair of symbols if they occur adjacently to each other in any of the hypothetical root words. For example, if the triplet of symbols σ_i σ_j σ_k is a hypothetical root word, then we connect the vertices for σ_i and σ_j and also connect the vertices for σ_j and σ_k.
4. **Test that the graph has at least two major connected components**. Vowel harmony can be detected by the presence of two large connected components in the graph. The first connected component consists of syllables with only back vowels, while the second connected component consists of syllables with only front vowels. Since languages have many consonants and several back and front vowels, the main connected components are expected to be large and about equal in size. Note that an isolated pair of vertices would mean that some word is composed of rarely used syllables. Such isolated pair of vertices may indicate either that we do not have enough documents in the library or the word is a borrowed word from another language that has different syllabic phonetics than the native language. In either case, the isolated pairs can be ignored.
5. **Return an answer**. Return 'vowel harmony found' if there are two large connected components. Otherwise, return 'language lacks vowel harmony.'

A. Applying the Vowel Harmony Testing Algorithm to the Phaistos Disk

Next we show as an example the application of the *Vowel-Harmony-Check*(λ, σ, τ) algorithm assuming that the library λ = Phaistos Disk, the script σ = the 45 Phaistos Disk symbols, and the threshold is τ = 2.

Step (2) of the *Vowel-Harmony-Check* algorithm identifies the fourteen hypothetical root words in the third column of Table 1 because all those pairs and triplets of symbols occur at least twice, and they do not always occur at the beginning or end of the blocks. The table displays an entire block of symbols in each row, read from left to right, for convenience. Note that the frequent pair of symbols  and  , which always occur at the beginning, and  , which always occurs at the end of blocks, are likely to be prefixes and suffixes, respectively, rather than roots. Hence they are not identified as hypothetical roots.

Step (3) of the algorithm creates the undirected graph shown in Fig. 1.

Step (4) of the algorithm would find the blue and the red connected components of the syllable adjacency graph as shown in Fig. 1. Both of these components have seven vertices, which is exactly what we are looking for, that is, two relatively big connected components of about equal size. The smaller connected components are all isolated pairs of vertices and can be ignored.

Step (5) of the algorithm would return 'vowel harmony found.'

TABLE I Phaistos Disk Blocks with Hypothetical Roots

Block	Prefix	Root	Suffix
A2			
A6			
A31			
B23			
B24			
A21			
B30			
A26			
A29			
B7			
B29			
B10			
B13			
A8			
A24			
A12			
B6			
B25			
B18			
B26			
B2			
B9			
A12			
B19			
A7			
A10			
A23			
B15			
B24			
B3			
B15			
A5			
A22			
A30			
B21			

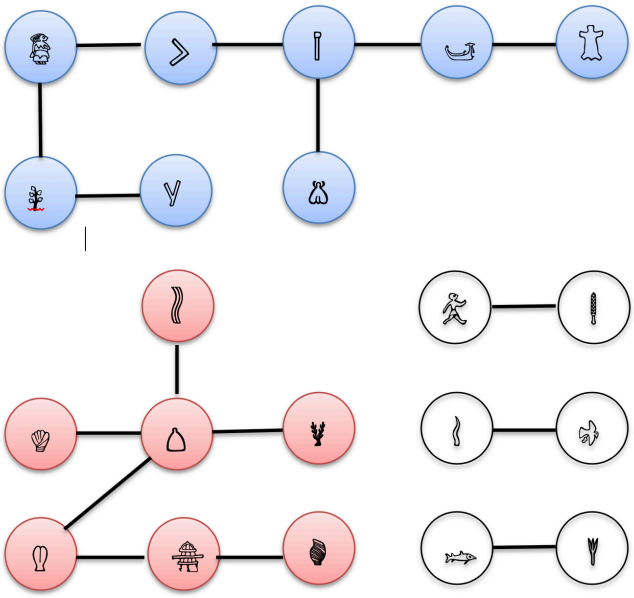

Figure 1: Syllable adjacency graph of the Phaistos Disk.

Block #	Phaistos Disk	Cretan Hieroglyphic	CHIC #
A1			260, 276, 298
A6			113a
A9			089b
A12			053aA
A23			320
B2			127, 171
B2			303a
B4			113b1
B13			060c
B17			125
B23			189

Figure 2: Matches between the Phaistos Disk and Cretan Hieroglyphic inscriptions.

III. Considerations of Cretan Hieroglyphs

Although the Phaistos Disk is considered by a minority of authors to be separate from the Cretan Hieroglyphs, Fig. 2 shows that many sign sequences within the Phaistos Disk appear to match Cretan Hieroglyph sign sequences. Hence that gives a confidence that they are simply variant inscriptions of the same Minoan language.

One very interesting sign sequence occurs in block A1 of the Phaistos Disk, which has three similar sequences among the Cretan Hieroglyph inscriptions [11], including on seals that contain only this sequence and nothing else. That suggests that this sequence likely denotes a root word. Therefore, the three signs:

, and

which we did not see yet as part of root words, now can be added to the syllable adjacency graph as another connected component.

Similarly, the sequence also seems to be a root word because it occurs just before the frequently occurring sequence, which we argued is likely to be a suffix. If the match is correct with the Cretan Hieroglyph seal, then now there is a second occurrence of this sequence, and we can also add the sign to the syllable adjacency graph. It would be part of the blue connected component.

We are exploring using the algorithm to automatically find additional matches between the Phaistos Disk and the Cretan Hieroglyph inscriptions until all the Phaistos Disk signs eventually can be joined either to the blue or the red connected component in the syllable adjacency graph. One of the difficulties in the process is that in general the Cretan Hieroglyph inscriptions do not indicate as clearly the word or at least phrase divisions as the Phaistos Disk does by its block structure.

IV. Discussion of the Algorithm

A. Reading Direction Independence

The first observation about the algorithm is that it can be applied even if we do not know the correct reading direction. That is possible because the prefixes and suffixes would be interchanged, but the root words would be identified similarly. Although all the affixes and the root words would be read in a reverse order, the adjacent syllables would remain adjacent. Hence the syllable adjacency graph would remain the same and the conclusions would be made regarding the number and sizes of the connected components. This observation is important because one of the perennial problems with ancient inscriptions is that the reading direction is often uncertain.

The second observation is that the identification of roots is easiest when the language is agglutinative, although they may also be identified in other types of languages. In Table 1 there are several symbols that are likely to be affixes, because they occur with high frequencies before or after the root words. For example, 𐄁, 𐄐 and 𐄑 occur four times each as apparent prefixes or suffixes. In some languages the vowel harmony extends to the prefixes and suffixes. Here 𐄁 and 𐄐 never occur before the blue group of signs but always with the red group of signs and the 𐄒 𐄓 isolate group. Some isolate groups may join the blue or the red groups if more inscriptions are studied. Hence this distribution of 𐄁 and 𐄐 suggests that in the Minoan language the vowel harmony extends to these apparent prefixes.

B. Implications for the Translation of the Minoan Inscriptions

Our algorithm shows that the Minoan language has vowel harmony. Hence, proposed translations of Minoan inscriptions that assume that the underlying Minoan language is Basque [9], an Indo-European language such as Greek [5, 6, 14]), a Kartvelian language [11] or a Semitic language [2, 10] now can be proven to be misguided. However, based on the assumption that Minoan is a Finno-Ugric language, Revesz presented translations of the Phaistos Disk [17], several Cretan Hieroglyphic inscriptions [18, 19] and over twenty Linear A inscriptions [20]. These translations remain viable proposals in light of the fact that Finno-Ugric languages have vowel harmony.

V. Conclusions and Future Work

The great advantage of the vowel harmony testing algorithm is that it can be applied to any syllabic script. We will present an application of the vowel testing algorithm to Linear A [8, 16], the Indus Valley Script [3] and Sumerian [21] inscriptions in an extended journal version of this paper. The result of the vowel harmony test can help categorize the underlying language as one that either contains or does not contain vowel harmony. This categorization helps to narrow down the possible underlying languages that need to be considered in decipherment algorithms such as the method of Snyder et al. [22].

References

[1] K. Aartun, Der Diskos von Phaistos; Die beschriftete Bronzeaxt; Die Inschrift der Taragona-tafel in Die minoische Schrift : Sprache und Texte vol. 1, Wiesbaden, Harrassowitz, 1992.

[2] W. Achterberg, J. Best, K. Enzler, L. Rietveld, and F. Woudhuizen, The Phaistos Disc: A Luwian Letter to Nestor, Publications of the Henry Frankfort Foundation vol XIII, Dutch Archeological and Historical Society, Amsterdam, 2004.

[3] S. Daggumati, and P. Z. Revesz, "Data mining ancient scripts to investigate their relationships and origins," Proceedings of the 23rd International Database Engineering and Applications Symposium, ACM Press, pp. 209-218, 2019.

[4] Y. Duhoux, 1998, "Pre-Hellenic Language(s) of Crete," The Journal of Indo-European Studies, Vol. 26 (1-2), pp. 1-38, 1998.

[5] J. Faucounau, Le déchiffrement du Disque de Phaistos & Les Proto-Ioniens: Histoire d'un peuple oublié, Paris, 1999.

[6] S. R. Fischer, Evidence for Hellenic Dialect in the Phaistos Disk, Herbert Lang, 1988.

[7] V. Georgiev, 1976, "Le déchiffrement du texte sur le disque de Phaistos." Linguistique Balkanique. vol. 19, pp. 5–47.

[8] L. Godart and J.-P. Olivier, Recueil des inscriptions en Linéaire A (Études Crétoises 21), Paris: De Boccard, 1976.

[9] F. G. Gordon, Through Basque to Minoan: transliterations and translations of the Minoan tablets. London: Oxford University Press, 1931.

[10] C. H. Gordon, Evidence for the Minoan Language, Ventnor, NJ: Ventnor Publishing, 1966.

[11] G. Kvashilava, On the Phaistos Disk as a Sample of Colchian Goldscript and Its Related Scripts, available on the author's webpage, 2008.

[12] I. Kenesei, R. M. Vago, and A. Fenyvesi. Hungarian. Routledge, 2002.

[13] K. Massey and K. Massey, Is the Phaistos Disk Cracked?, 1998.

[14] G. Nagy, "Greek-Like Elements in Linear A," Greek, Roman, and Byzantine Studies, Harvard University Press , vol. 4, pp. 181–211, 1963.

[15] J.-P. Olivier, L. Godart and J.-C. Poursat, Corpus Hieroglyphicarum Inscriptionum Cretae (Études Crétoises 31), De Boccard, 1996.

[16] T. Petrolito, et al. "Minoan Linguistic Resources: The Linear a Digital Corpus." Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics and The Asian Federation of Natural Language Processing, 2015.

[17] P. Z. Revesz, "A computer-aided translation of the Phaistos Disk," International Journal of Computers, vol. 10, pp. 94–100, 2016.

[18] P. Z. Revesz, "A computer-aided translation of the Cretan Hieroglyph Script," Inter. Journal of Signal Processing, vol. 1, pp. 127-133, 2016.

[19] P. Z. Revesz, "A translation of the Arkalochori Axe and the Malia Altar Stone," WSEAS Transactions on Information Science. and Application, vol. 14, pp. 124-133, 2017.

[20] P. Z. Revesz, "Establishing the West-Ugric language family with Minoan, Hattic and Hungarian by a decipherment of Linear A," WSEAS Trans. on Information Science and Applications, vol. 14, pp. 306-335, 2017.

[21] P. Z. Revesz, "Sumerian contains Dravidian and Uralic substrates associated with the Emegir and Emesal dialects," WSEAS Transactions on Information Science and Applications, vol. 16, pp. 8-30, 2019.

[22] B. Snyder, R. Barzilay, and K. Knight, "A Statistical Model for Lost Language Decipherment," Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1048-1057, 2010.