Peer-reviewed version available at *Math. Comput. Appl.* **2019**, *24*, 14; doi:10.3390/mca24010014

Article

A Hidden Markov Model for the Linguistic Analysis of the Voynich's Manuscript

L. Acedo

Instituto Universitario de Matemática Multidisciplinar
Building 8G, 2º Floor, Camino de Vera,
Universitat Politècnica de València, 46022, Valencia, Spain.; luiacrod@imm.upv.es
* Correspondence: luiacrod@imm.upv.es; Tel.: +34-963877007 Ext: 88285

Abstract: Hidden markov models are a very useful tool in the modelling of time series and any sequence of data. In particular, they have been successfully applied to the field of mathematical linguistics. In this paper, we apply a hidden markov model to analyze the underlying structure of an ancient and complex manuscript, known as Voynich's manuscript, that remains still undeciphered. By assuming a certain number of internal states representations for the symbols of the manuscript we train the network by means of the α and β-pass algorithms to optimize the model. By this procedure, we are able to obtain the so-called transition and observation matrices in order to compare with known languages concerning the frequency of consonant and vowel sounds. From this analysis, we conclude that transitions occur between the two states with similar frequencies to other languages. Moreover, the identification of the vowel and consonant sounds matches some previous tentative bottom-up approaches to decode the manuscript.

Keywords: Hidden Markov Models; Mathematical Linguistics; Voynich Manuscript

1. Introduction

Hidden Markov models (HMMs) are a particular kind of a Bayesian network obtained by combining a Hidden Markov layer and a second layer of outputs that depends probabilistically on the hidden states of the first layer [1,2]. This model has proven a very versatile and useful tool in many applications of artificial intelligence including: (i) modelling of biological sequences of proteins and DNA [3] (ii) speech recognition systems [4] (iii) data compression and pattern recognition [5] (iv) object tracking in video sequences [6] and other. Of particular interest to us are the early studies in which HMMs were used to analyze a large body of text, in this case of English (the so-called "Brown Corpus"), considered as a sequence of letters without any previous assumption on the linguistic structure of the text or the meaning of the letters [1,4,7]. Depending on the number of hidden states, thanks to this work light was shed on the linguistic structure of English depending on the number of hidden states considered in the model. For example, for two hidden states the basic division among vowels and consonants was recovered as the most natural basic structure of English language [7]. As many more states were taken into account it was discovered a structure including the initial and final letters of a word, vowel followers and preceders, etc. This elucidates the purely statistical nature of a language and it shows that HMMs can be an insightful tool in mathematical and computational linguistics.

Applications of HMMs to the emergent field of Natural Language Processing (NLP) has also flourished in recent years as it has been shown its applicability to different layers of NLP such as speech tagging and morphological analysis. By using this approach successful results for many languages such as Arabic and Persian have been obtained [8,9]. For these reasons, it seems promising to extend these analyses to other sources of text that cannot still be deciphered because they are written in an unknown script and with a unique linguistic structure. Among the candidates to this challenge, it stands out the medieval codex known as Voynich's manuscript [10].

Discovered by the Polish-Samotigian book dealer W. Voynich in 1912, it have remained an enigma of historical cryptography since then. For a detailed introduction to the manuscript history and attempts of decipherment until 1978 the interested reader can found more information in M. d'Imperio monograph [11] and Zandbergen's website [10]. It is common lore about history researchers that this book could have belonged to emperor Rudolph II of Baviera until his death in 1612 as stated in a letter addressed to the XVIIth century scholar Athanasius Kircher that was found by Voynich himself in the manuscript. A reconstruction of the history of the ownership of the manuscript has been elucidated throughout the years. It is also known that it was property of the Jesuits and it was kept at the "Collegio Romano" since the last decade of the XVIIth century until the end of the XIXth century when it was moved to Frascati where Voynich acquired it.

Modern physics and chemistry analyses have allowed establishing some rigorous facts. Firstly, in 2009 some samples of ink and paint were taken from the manuscripts and analysed by X-ray spectroscopy and X-ray diffraction techniques showing that these inks and pigments were totally compatible with those used by scribes at the last epoch of the middle ages [12]. The same year a radiocarbon dating of the parchment was carried out by researchers at the University of Arizona [13]. They found that with a 95% probability the parchment corresponds to the period between 1404 and 1438 placing the manuscript in the first half of the fifteen century. It is also clear that the text was added after the drawing of the figures in the manuscript because it usually surrounds the figures very closely.

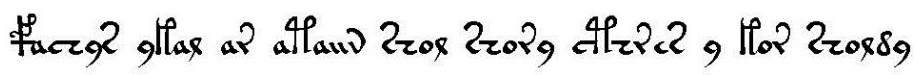


Figure 1. A sample of the first folio of the Voynich's manuscript.

An image of the first line in the Voynich's manuscript is shown in Fig. 1. The total set of individual characters depends on some ambiguities in counting but it seems that there are 36 characters in the set as recognized by the European Voynich Alphabet (EVA), some of them far more frequent than other. Although the symbols seems strange, and not immediately associated with any known alphabet ancient or modern, a closer inspection reveals a similarity with Latin, Arabic numerals and, specially, some Latin abbreviations very common throughout the middle ages [14]. Anyway, these clues have helped little in finding an accepted decipherment of the text.

Among the possible solutions to this riddle there have been four main proposals:

1. It is a manuscript written in an extinct natural language with an exotic alphabet [15].
2. It is the encipherment of a known language (possibly Latin, German or other Indo-european language but nobody is sure [11]).
3. It is a hoax consisting of asemic writing with the objective of making the book strange and valuable to collectors of antiquities [16].
4. It is a modern fabrication (perhaps by his discoverer, W. Voynich) [10].

From all these hypotheses, the last one seems excluded by modern physicochemical analyses but the other three may be considered still open. The paper is organized as follows: In Sec. 2 we discuss the basics of Hidden Markov Models and its application to linguistic analysis. Sec. 3 is devoted to the application of HMMs to the Voynich's manuscript and the information we may deduce from this. Finally, the paper ends with a discussion on the meaning of the findings of the paper and guidelines for future work in Sec. 4.

2. Hidden Markov models

In this section we will provide a quick summary of the basic concepts and algorithms for HMMs. In Fig. 2, we show the structure of a HMM. The Markov process (above) is characterized by a sequence $\{X_0, X_1, X_2, \ldots, X_{T-1}\}$ of internal states selected among a total of N. The transition among these states is performed according to the probabilities in a transition matrix $\boldsymbol{A}$ in such a way that the element

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2018 doi:10.20944/preprints201811.0594.v1

Peer-reviewed version available at Math. Comput. Appl. 2019, 24, 14; doi:10.3390/mca24010014

a_{ij} denotes the probability of performing a transition from the internal i state to the j state. We can also denote the different internal states as q_i, with $i = 0, 1, 2, \ldots, q_{N-1}$.

The second layer we have plotted in Fig. 2 corresponds to the observations. The sequence of observations is then denoted by $\{\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_{T-1}\}$ and they can be chosen among a total of M possible observation symbols. The relation between the Markov process' layer and the observation layer is also probabilistic because, given an internal state q_j, the probability for observing the symbol k is $b_j(k)$. These elements constitute a row stochastic matrix, $\boldsymbol{B}$.

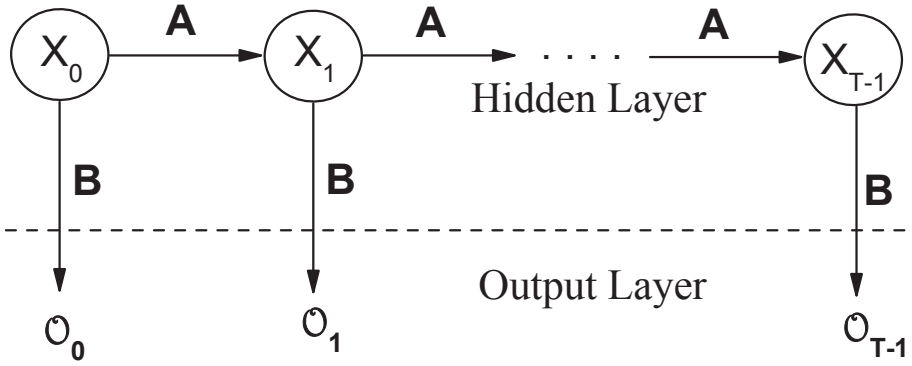


Figure 2. A schematic view of a HMM. See the main text for details.

The problem we want to address is the following: Suppose we have a given sequence of observations, $\mathcal{O}$, consisting on a series of symbols from a total of M. If we assume that there are N internal states in the model, the objective is to find the model $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ (where $\boldsymbol{\pi}$ is the distribution of initial states) that provides the best fit of the observation data. The standard technique in HMMs to evaluate this optimum model make use of a forward and a backward algorithm described as follows.

2.1. The forward algorithm

Firstly, we are interested in evaluating the following probability:

$$\alpha_t(i) = P(\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_t, x_t = q_i|\lambda) \,, \tag{1}$$

i. e., the probability that the sequence of observations up to time t is given by $\mathcal{O}_0, \ldots, \mathcal{O}_t$ and the internal state at time t is q_i for a given model λ. Then, for $t = 0$ we have that:

$$\alpha_0(i) = \boldsymbol{\pi}_i \, b_i(\mathcal{O}_0) \,, \tag{2}$$

where $i = 0, 1, \ldots, N-1$. The reason is that $\boldsymbol{\pi}_i$ is the probability that the initial internal state is q_i, and $b_i(\mathcal{O}_0)$ is the probability that, given that the internal state is q_i, we have the observation $\mathcal{O}_0$. It is the easy to check that the recursion expression for $\alpha_t(i)$ is given by:

$$\alpha_t(i) = \left[\sum_{j=0}^{N-1} \alpha_{t-1}(j) \, a_{ji} \right] b_i(\mathcal{O}_t) \,. \tag{3}$$

Here a_{ji} is the transition probability from the inner state j to the inner state i. This algorithm is also called α-pass.

2.2. The backward algorithm

The backward algorithm or β-pass is proposed for the efficient evaluation of the following probability:

$$\beta_t(i) = P(\mathcal{O}_{t+1}, \mathcal{O}_{t+2}, \ldots, \mathcal{O}_{T-1}, x_t = q_i|\lambda) \,. \tag{4}$$

This means that we interested in finding the probability that the sequence of observations from time $t+1$ to the end is $\mathcal{O}_{t+1}, \ldots, \mathcal{O}_{T-1}$ and the inner state at time t is q_i. The algorithm is constructed as follows:

- In the first place, we define $\beta_{T-1}(i) = 1$ for $i = 0, 1, \ldots, N-1$.
- Then, for $t = T-2, T-3, \ldots, 0$ we define the recursive relation:

$$\beta_t(i) = \sum_{j=0}^{N-1} a_{ij}\, b_j(\mathcal{O}_{t+1})\, \beta_{t+1}(j) \,. \tag{5}$$

We will use these two algorithms, the forward and the backward, to find a new algorithm that can be used to re-estimate the model and make it approach to its optimum.

2.3. Reestimating the model

Our objective now is to reevaluate the model in such a way that the optimum parameters that fit the observations are found. These parameters are the elements of the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and also those in the vector corresponding to the initial distribution of internal states, $\boldsymbol{\pi}$. We need now an algorithm to reestimate the model in such a way that the probability of the observation sequence, $\mathcal{O}_0, \ldots, \mathcal{O}_{T-1}$, given the model λ, $P(\mathcal{O}_0, \ldots, \mathcal{O}_{T-1}|\lambda)$ is maximized.

The idea of the algorithm begins with the definition of the following probability for a given model and a given observation sequence:

$$\gamma_t(i,j) = P(x_t = q_i, x_{t+1} = q_j | \mathcal{O}, \lambda) \,. \tag{6}$$

This is the probability of finding the internal states q_i, q_j at times t, $t+1$ for the observation sequence $\mathcal{O}$ and the model λ. Using now the standard relations for conditional probabilities and the definitions of the α and β probabilities in Eqs. (1), (4) we have:

$$\gamma_t(i,j) = \frac{\alpha_t(i)\, a_{ij}\, b_j(\mathcal{O}_{t+1})\, \beta_{t+1}(j)}{P(\mathcal{O}|\lambda)} \,, \tag{7}$$

for time $t = 0, 1, \ldots, T-2$. We also define the sum over the index j, i. e., the probability of finding the inner state q_i at time t for a given model and observation sequence:

$$\gamma_t(i) = \sum_{j=0}^{N-1} \gamma_t(i,j) \,. \tag{8}$$

With these definitions and expressions we can now propose the evolution algorithm for the reestimation of the parameters:

- We initialize the model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$. It is a common practice to choose the elements according to the uniform distribution: $\boldsymbol{\pi}_i \approx 1/N$, $a_{ij} \approx 1/N$, and $b_j(k) \approx 1/M$ but these values must be randomized to avoid that the algorithm becomes stuck at a local maximum.
- We calculate the parameters $\alpha_t(i)$, $\beta_t(i)$, $\gamma_t(i,j)$ and $\gamma_t(i)$ by applying the corresponding expressions in Eqs. (3), (5), (7) and (8).
- For $i = 0, 1, \ldots, N-1$ and $j = 0, 1, \ldots, N-1$ we reestimate the elements of the transition matrix, $\boldsymbol{A}$, as follows:

$$a_{ij} = \frac{\sum_{t=0}^{T-2} \gamma_t(i,j)}{\sum_{t=0}^{T-2} \gamma_t(i)} \,, \tag{9}$$

Peer-reviewed version available at Math. Comput. Appl. 2019, 24, 14; doi:10.3390/mca24010014

- For $i = 0, 1, \ldots, N-1$ and $j = 0, 1, \ldots, N-1$ we compute the new values for the elements of the observation probability matrix as follows:

$$b_j(k) = \frac{\sum_{t\in\{0,1,\ldots,T-1\}\,,\,\mathcal{O}_t=k} \gamma_t(j)}{\sum_{t=0}^{T-1} \gamma_t(j)} \,. \tag{10}$$

  Here the sum in the numerator is restricted to those instants of time in which the observation symbol is the k-th.
- Finally, we compute the probability of the given observation sequence, i.e., $P(\mathcal{O}|\lambda)$ (obtained as the sum of $\alpha_{T-1}(i)$ for all the inner state values, i). If this probability increases, with respect to the previous value, the model updating is performed again. However, in practice the algorithm is run for a given number of steps or until the probability does not increases more than a selected tolerance.

Another issue with this algorithm is that the α-pass and β-pass evaluations may easily lead to underflow. To avoid this problem, a normalization by the sum over j of $\alpha_t(j)$ is performed. For further details on HMMs the interested reader may check the references [1].

3. Results

In this section we will discuss the application of the algorithm discussed in the previous section to several cases. First, we will consider the case of a text in English and we will implement the model optimization algorithm to classify the letters of the alphabet (after removing all the punctuation signs) into two classes corresponding to the inner states of the HMM. It will be shown that these classes are clearly associated with the vowels and the consonants in English and this provides the basic phonemic structure of the language. Testing the algorithm with a known language gives us the necessary confidence to apply it to the Voynich's manuscript. This way we will show that both vowel and consonant sounds are identified in the manuscript with a small subset of ambiguous symbols that may function as both a vowel or a consonant. This may be explained by implicit vowelling of some consonants, as it is commonly found in Abjad and Abugidas scripts.

3.1. Application to The Quixote:

In this section we will apply the model evolution algorithm for HMM with $N = 2$ and $M = 27$. Therefore, we consider a total of 26 letters and the space as output symbols. The text of the Quixote in plain ASCII can be freely downloaded from the Gutenberg's project website [**?**]. As a data pre-processing stage we transform all the upper-case letters to lower-case and remove the punctuation signs with the exception of the spaces among subsequent words. This way we obtain a sequence of $5,693,310$ characters but, to our purpose, we can restrict ourselves to the first $100,000$ characters.

As initial transition matrix we have chosen:

$$\boldsymbol{A} = \begin{pmatrix} 0.46 & 0.54 \\ 0.52 & 0.48 \end{pmatrix} \,, \tag{11}$$

and the distribution of initial hidden states is given at the start of the algorithm by

$$\boldsymbol{\pi} = \begin{pmatrix} 0.52 \\ 0.48 \end{pmatrix} \,, \tag{12}$$

The observation probability matrix is obtained by randomizing the equal probability assumption: $b_j(\mathcal{O}) = 1/M$ for every j. For example, we can multiply $1/M$ by a random number in the interval

Peer-reviewed version available at Math. Comput. Appl. 2019, 24, 14; doi:10.3390/mca24010014

$(0.8, 1.2)$. Anyway, we must verify the condition that the total probability for a given inner state j and all the possible observation outcomes is normalized to one:

$$\sum_{\text{all } \mathcal{O}} b_j(\mathcal{O}) = 1 \,, \tag{13}$$

and this is accomplished by imposing this condition to define the last value $b_j(\mathcal{O})$ for the M-th state, $\mathcal{O}$.

The algorithm was then implemented in Mathematica using lists and the model evolution was run for 200 steps with $100,000$ characters from the book, pre-processed to retain only the letters and removing all the punctuation signs. Results were also checked using other independent implementations of the code in C++ [18].

In the first place, we notice that the algorithm converges very fast as deduced from the evolution of the logarithm of the observation sequence probability for the given model $P(\mathcal{O}|\lambda)$. As shown in Fig. 3

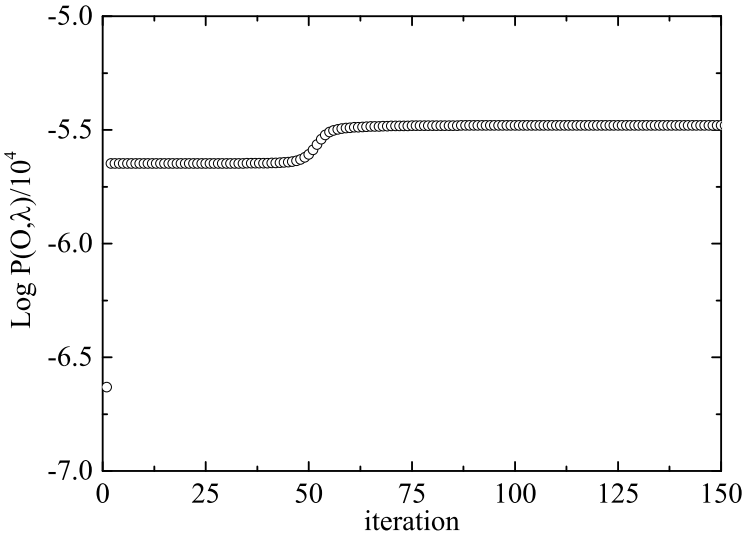


Figure 3. The evolution of the logarithm of the observation sequence's probability as a function of the iteration. Notice the fast convergence to an asymptotic "plateau".

From this figure we conclude that convergence is achieved after only 60 steps and that further iterations only improve the results very slightly. The final transition matrix we find is given as follows:

$$\boldsymbol{A} = \begin{pmatrix} 0.369 & 0.631 \\ 0.869 & 0.131 \end{pmatrix} \,, \tag{14}$$

with an initial distribution of initial states given by:

$$\boldsymbol{\pi} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \,. \tag{15}$$

This is also found using other texts such as the "Brown Corpus". The form of the transition matrix in Eq. (14) is already pointing towards the existence of two categories of letters (the inner state of the HMM). These categories are, obviously, the vowels and the consonants and this statement is reinforced

Peer-reviewed version available at Math. Comput. Appl. 2019, 24, 14; doi:10.3390/mca24010014

by the values of the observation probability matrix, $\boldsymbol{B}$. In our case, the inner state 2 has been identified as the vowels as it is clearly shown in Fig. 4 where a peak of probability is found for every vowel.

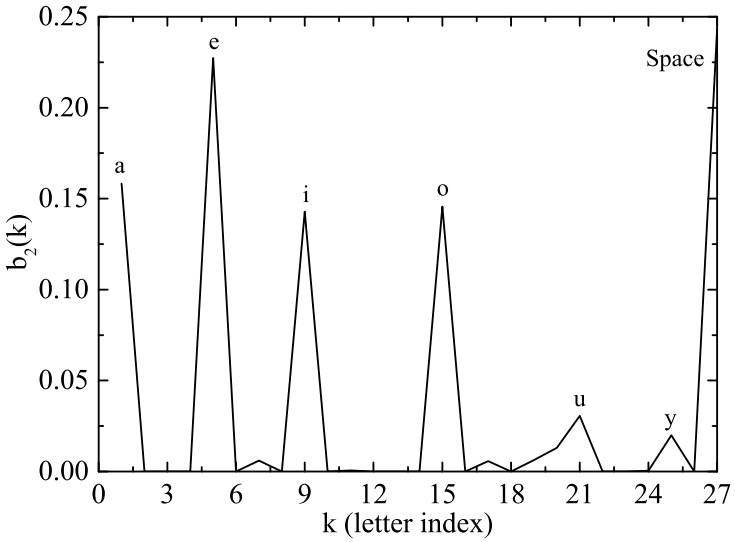


Figure 4. The probability of observation of a given letter for the hidden state 2. Notice the peaks for the vowels as well as "y" and the space.

From Fig. 4 we can also derive some interesting conclusions:

1. The most frequent vowel in the English language is "e".
2. The space among words has the structural function of a vowel although it has no associated sound.
3. The letter "y" is mostly a vowel in the English language. Indeed, the Oxford Dictionary classifies it as a vowel in some cases ("myth"), a semivowel in others ("yes") or as a part of a diphthong (as in "my") [19].

Similar conclusions can be also deduced for the consonants. It is also remarkable that the results in Fig. 4 cannot be deduced from simple frequency analysis. The histogram of the letters obtained from the Quixote's text is shown in Fig. 5. We see that the vowel "e" is still the most frequent letter in the English language (not only the most frequent vowel) but the second most frequent letter is "t". So, there is no clear pattern in the histogram to separate vowels from consonants. This ability of HMM make them a very powerful tool in computational linguistics.

3.2. Application to the Voynich's manuscript

After the successful implementation of the HMM technique for the Quixote we now turn again to our problem of analyzing the Voynich's manuscript. Several transcriptions of this manuscript are available but the most popular is the one based upon the so-called European Voynich Alphabet (EVA) as developed by R. Zandbergen and G. Landini. Although there is an extended version which includes the less common symbols in the Voynich's manuscript, the basic version uses 26 letters of the English alphabet (excluding "w") to make a correspondence with the most abundant symbols in Voynichese. The correspondence among the EVA code and the Voynich's symbol is given in the table of Fig. 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2018 doi:10.20944/preprints201811.0594.v1

Peer-reviewed version available at Math. Comput. Appl. 2019, 24, 14; doi:10.3390/mca24010014

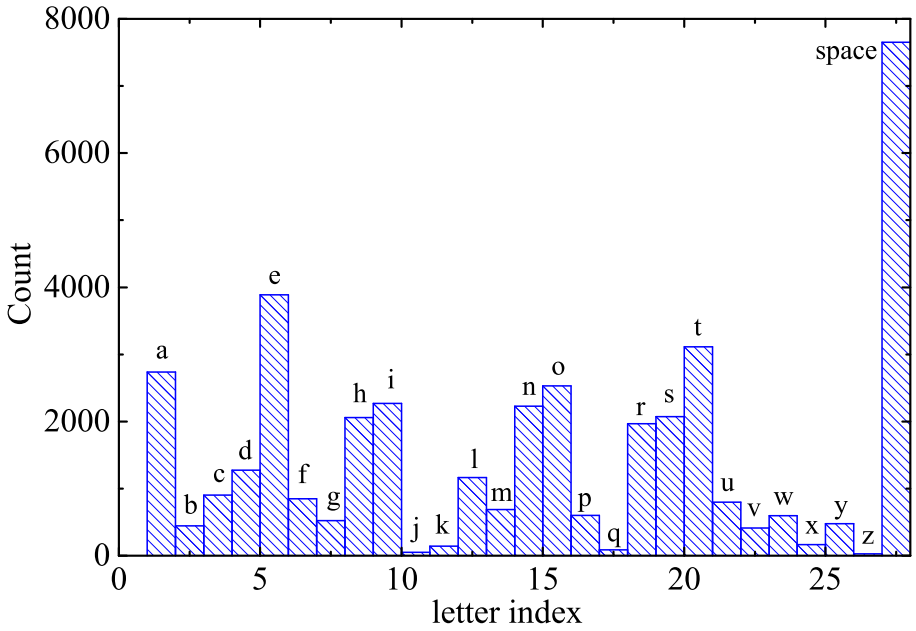


Figure 5. The histogram for the frequency of the totality of letters (and the space) in the English version of The Quixote.

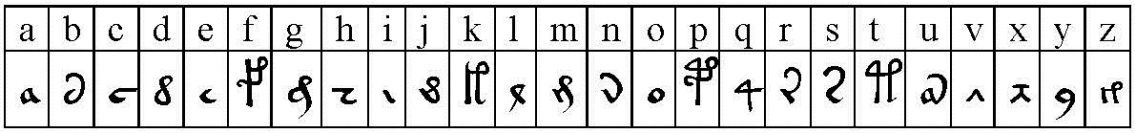


Figure 6. The correspondence among Voynich's symbols and the associated letter in the EVA transcription systems. Notice that this is merely an arbitrary codification without any relation to the actual phonemes that the Voynich's characters might represent.

This way a transcription of the whole Voynich's manuscript has been performed in such a way that it can be used in computational analysis. These transcriptions are available in several sites and they are discussed by Zandbergen [10]. In particular, we will use the Takahashi's transcription developed in 1999. Of course, some pre-processing is required before applying the HMM algorithm because this file includes some information about each line, including the folium number (recto or verso) and the number of the line within each page of the manuscript. After removing this information, we are left with a set of EVA characters separated by dots. These dots correspond to the spaces between words in the original manuscript. The total of characters (including spaces) is 228,836 that we can use for the simulation. Convergence of the algorithm was also very fast, as in the case of the English text analyzed in Sec. 3.1, and if we use more than 50,000 characters the results are stable and show no dependence on the total length of the sequence, T. This is a convincing argument in favour of the consistency of the results.

We started with the same initial conditions as those given in Eqs. (11) and (12) for the transition matrix and the distribution of the state $t = 0$. The probability matrix for the observation states (the Voynich characters) was randomized in the usual way explained in Sec. 3.1. In this particular example we used the first 100,000 characters in the Voynich's manuscript and 200 iteration steps. The final transition matrix shows the same pattern as those of natural languages:

$$\boldsymbol{A} = \begin{pmatrix} 0.169 & 0.831 \\ 0.840 & 0.160 \end{pmatrix}, \tag{16}$$

Peer-reviewed version available at Math. Comput. Appl. 2019, 24, 14; doi:10.3390/mca24010014

and the distribution of initial states is given by:

$$\pi = \begin{pmatrix} 0 \\ 1 \end{pmatrix} . \tag{17}$$

We will see that this indicates that the first character of the manuscript is associated with a consonant sound. The most interesting results are, however, those obtained with the observation probability matrix, which clearly separate two kinds of characters to be associated with vowel and consonant phonemes as it occurs with the case of The Quixote. In Fig. 7, we show the probability of obtaining each character when the hidden state is 1.

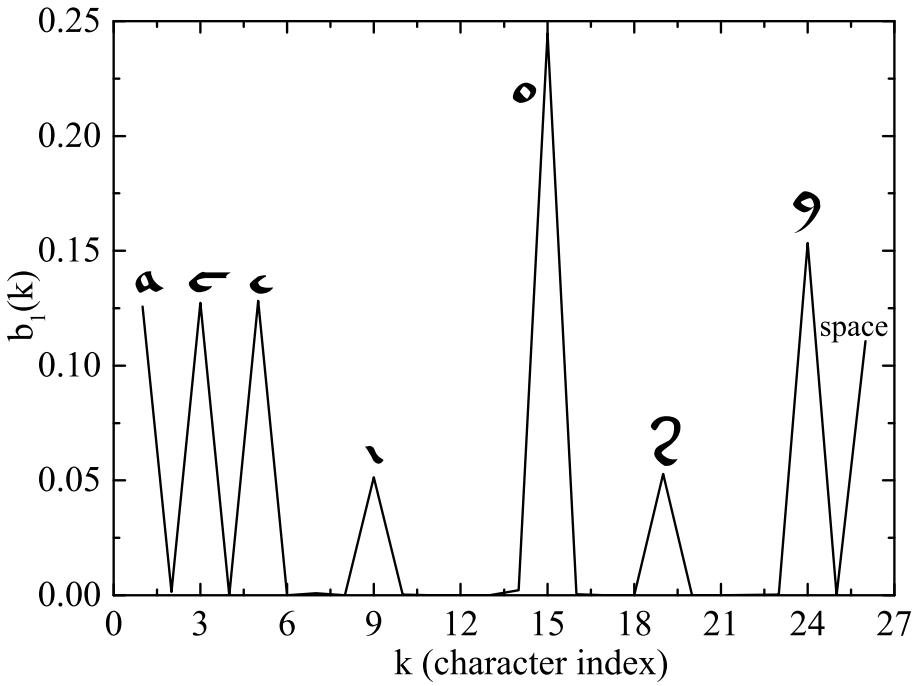


Figure 7. The probability of observation of a given character in the Voynich's manuscript for the hidden state 1. The peaks corresponds to the symbols: "a", "c", "e", "i", "o" and "y" of the EVA alphabet. There is also a peak for the space among words.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2018 doi:10.20944/preprints201811.0594.v1

Peer-reviewed version available at Math. Comput. Appl. 2019, 24, 14; doi:10.3390/mca24010014

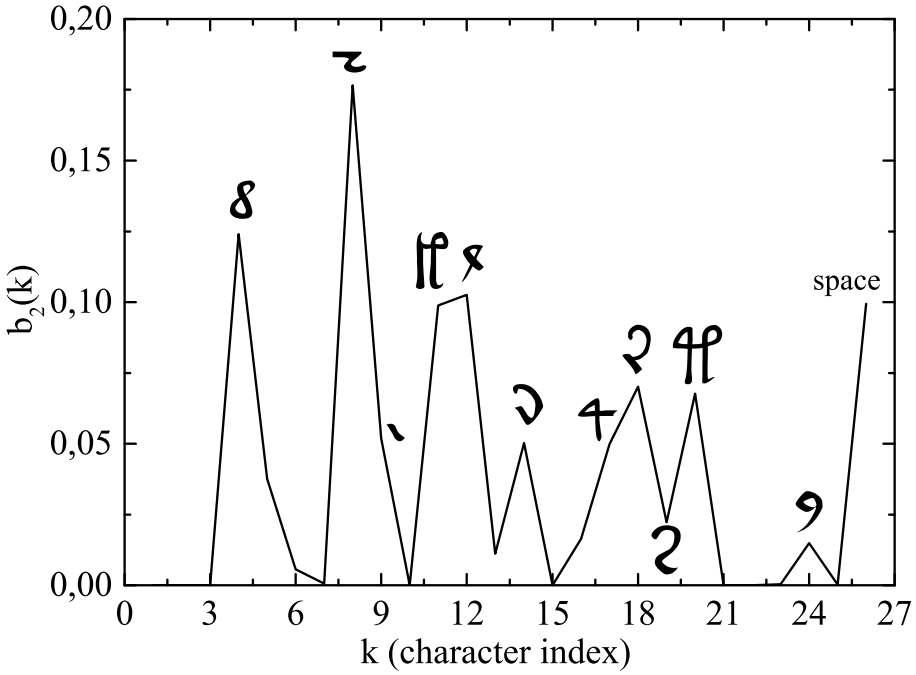


Figure 8. The same as figure 7 but for the hidden state 2.

The probabilities for hidden state 2 are given in Fig. 8. We see that a set of very conspicuous peaks are obtained in both cases but there are fewer in Fig. 7, this could mean that the symbols corresponding to those peaks are associated with vowel's phonemes. On the other hand, this correspondence is not as strong as in the case of the English text of Sec. 3.1 because there are symbols with noticeable probability, that appear in both figures (in particular, the EVA symbols "i", "s" and "y"). The fact that the separation among vowels and consonants is not as clear as in the case of European languages may imply that we are dealing with some implicit vowelling in the language underlying the manuscript as it is common in Abjads' alphabets such as that of Arabic. Nevertheless, the appeareance of high peaks in Fig. 7 also suggests the existence of vowel symbols in the manuscript so this Abjad would be impure. Bax also suggested in 2014 [15] that this is the case after proposing some provisional decoding of some words in the manuscripts associated with plants and a constellation. In the next section, we will discuss this intriguing possibility.

4. Discussion

Research into the Voynich's manuscript has been dominated by completely erred perspectives since its rediscovery in the early XXth century [11]. Most of these approaches have been unscientific and very subjective and they have hindered the establishment of a serious research project leading to the understanding of this unique book. The recent proof of the old age of the "vellum" by carbon-dating places this manuscript in the first half of the XVth century and the rest of tests are very compatible with an origin in the middle ages [12,13]. Anyway, the situation is still so confusing that some authors still support the idea that this object was fabricated with a pecuniary intention by making it similar in appearance to a real, but enciphered, text [16]. On the other hand, statistical measurements carried out by Amancio et al. in 2013 [20] show that the Voynich's manuscript is incompatible with shuffled texts and, moreover, that certain keywords appear throughout the manuscript. These keywords organize in patterns of semantic networks as shown by Montemurro et al. [21].

Other authors, as most researchers in the past, think that the text is enciphered in some way. Hauer and Kondrak assumed that the manuscript is written in some Abjad's alphabet using some transposition of letters or anagramming [22]. These assumptions are very strong and their conclusions of a relation to Hebrew have been widely dismissed.

In this paper, we use the conservative approach of applying the standard technique of HMM for the linguistic analysis of the manuscript. We have shown that a division among vowels and consonant phonemes is very clear in the resulting observation probability matrix but that some characters (such as the EVA symbols "i", "s" and "y") could participate of the vowel and consonant nature, either because they are semivowels or because there is an implicit vowelling in them as in impure Abjad's alphabets. A positional dependence of the vowelling as in Abugida's alphabets cannot be excluded. In any case, we found some characters in the script that very possibly represent vowel's phonemes: these are "a", "c", "e" and "o" in the EVA notation (See Fig. 6). There are also some exclusively consonant phonemes such as "d", "h", "k", "l" or "t" in the EVA's notation.

It seems interesting that Prof. Bax in 2014 [15] identified some names of plants and the constellation Taurus in the manuscript and that his associations with phonemes are similar to the one discussed in this paper. In Fig. 9 we show some of these associations for the words "Taurus", "Coriander" and "Juniper" (in arabic).

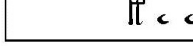
Voynich's Word	Phonetics' transcription
	/a/ r /a/ r
	T /a/ /ə/ /r/ (plus vowel) N
	K O O R A T /ə/ ?

Figure 9. Some words in the Voynich's manuscript and the phonetics transcriptions proposed by S. Bax [15].

Although these associations are considered very preliminary, even by its author, the similarity with our identification of vowels and consonants is striking. So, we can gain some confidence that a serious scholarship effort could enhance these identifications and provide a sure path for further research. Although the rest of possibilities might still be open, we have increasing support to the view that the text in the manuscript is not hoax nor an intentional cipher but a genuine language written in an unknown script. Notwithstanding this progress, we are still far from identifying the language because it could even be a dead tongue, for which the script was devised. HMM analysis, however, supports the idea that the language in the Voynich's manuscript has the characteristics of an impure Abjad pointing towards an eastern origin and not merely a cryptic form of an European language as it has been considered by many authors [11]. We hope that this work stimulates further research by expert linguists that could shed additional light into this ancient enigma.

This research received no external funding.

Acknowledgments: The paper is dedicated to the late Prof. Stephen Bax whose enthusiastic work on the Voynich's manuscript has stimulated the research into this lingering enigma. The author also gratefully acknowledges Prof. Mark Stamp for providing the C++ code implementation of the HMM algorithm. Dr. René Zandbergen is acknowledged for developing the EVA transcription of the Voynich's text.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EVA	European Voynich Alphabet
HMM	Hidden Markov Models
NLP	Natural Language Processing

References

1. Stamp, M. A Revealing Introduction to Hidden Markov Models. **2018**. Available online on: http://www.cs.sjsu.edu/faculty/stamp/RUA/HMM.pdf (Accessed: November, 22th, 2018).

2. Ghahramani, Z. An Introduction to Hidden Markov Models and Bayesian Networks. *Int. J. Pattern Recognition and Artificial Intelligence* **2001**, *15(1)*, 9-42, doi:10.1142/S0218001401000836.
3. Yoon, B. J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics* **2009**, *10*, 402-415, doi: 10.2174/138920209789177575.
4. Juang, B. H. and Rabiner, L. R. Hidden Markov Models for Speech Recognition. *Technometris* **1991**, *33(3)*, 251-272. doi: 10.2307/1268779.
5. Bicego, M., Castellani, U. and Murino, V. Using hidden Markov models and wavelets for face recognition. In *12th International Conference on Image Analysis and Processing, 2003.Proceedings.*, IEEE Computer Society, 2003. doi: 10.1109/ICIAP.2003.1234024.
6. Lefèvre, S., Bouton, E., Brouard, T. and Vincent, N. A new way to use Hidden Markov Models for object tracking in video sequences. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, Barcelona, Spain, 2003. ISBN: 0-7803-7750-8. doi: 10.1109/ICIP.2003.1247195.
7. Cave, R. L. and Neuwirth, L. P. Hidden Markov models for English. In *Hidden Markov Models for Speech*, IDA-CRD, Princeton, NJ, 1980. Available at: https://www.cs.sjsu.edu/~stamp/RUA/CaveNeuwirth/index.html (Accessed: November, 22th, 2018).
8. Suleiman, D., Awajan, A. and Al Etaiwi, W. The Use of Hidden Markov Model in Natural ARABIC Language Processing: a survey. *Procedia Computer Science* **2017**, *113*, 240-247. doi: 10.1016/j.procs.2017.08.363.
9. Okhovvat, M. and Bidgoli, B. M. A Hidden Markov Model for Persian Part-of-Speech Tagging. *Procedia Computer Science* **2011**, *3*, 977-981. doi: doi:10.1016/j.procs.2010.12.160.
10. Zandbergen, R. The Voynich Manuscript. Website at: http://www.voynich.nu.
11. D'Imperio, M. E. The Voynich Manuscript: An Elegant Enigma. National Security Agency, Central Security Service. Fort George G. Meade, Maryland, U. S. A., 1978.
12. Repp, K. Materials Analysis of the Voynich Manuscript. Available online at: https://beinecke.library.yale.edu/sites/default/files/voynich_analysis.pdf (Accessed: November, 22th, 2018).
13. Zandbergen, R. The Radio-Carbon Dating of the Voynich MS. **2016**. Available online at: http://www.voynich.nu/extra/carbon.html (Accessed: November, 22th, 2018).
14. Capelli, A. The Elements of Abbreviation in Medieval Latin Paleography. Translated by Heimann, D. and Kay, R. University of Kansas Libraries, 1982. (Translation of the original, Lexicon abbreviaturarum, published in 1899). Available online at: https://kuscholarworks.ku.edu/bitstream/handle/1808/1821/47cappelli.pdf.
15. Bax, S. A proposed partial decoding of the Voynich script. **2014**. Unpublished manuscript available online at: https://stephenbax.net/wp-content/uploads/2014/01/Voynich-a-provisional-partial-decoding-BAX.pdf.
16. Rugg, G. and Taylor, G. Hoaxing statistical features of the Voynich Manuscript. *Cryptologia* **2017**, *41(3)*, 247-268, doi: 10.1080/01611194.2016.1206753.
17. Gutenberg Project. The Quixote by Miguel de Cervantes Saavedra. Available at: http://www.gutenberg.org/ebooks/996. (Accessed: November, 21th, 2018).
18. An implementation in C++ of the HMM algorithm developed by M. Stamp is available at this website: http://www.cs.sjsu.edu/faculty/stamp/RUA/HMM_ref_fast.zip.
19. Oxford Dictionaries online. Is the letter "Y" a vowel or a consonant ?. https://en.oxforddictionaries.com/explore/is-the-letter-y-a-vowel-or-a-consonant/
20. Amancio, D. R., Altmann, E. G., Rybski, D., Oliveira Jr., O. N. and da F. Costa, L. Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript. *PLOS One* **2013**, *8(7)*, e67310. doi: 10.1371/journal.pone.0067310.
21. Montemurro, M. A. and Zanette, D. H. Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis. *PLOS One* **2013**, *8(6)*, e66344. doi: :10.1371/journal.pone.0066344.
22. Hauer, B. and Kondrak, G. Decoding Anagrammed Texts Written in an Unknown Language and Script. *Transactions of the Association for Computational Linguistics* **2016**, *4*, 75-86.