Deciphering Secrets Throughout History: An Interdisciplinary Linguistics and Cryptology Course

Eunice Kim
Classics Department
Furman University
Greenville, SC 29613
eunice.kim@furman.edu

Christian Millichap
Mathematics Department
Furman University
Greenville, SC 29613
christian.millichap@furman.edu

Abstract

This paper describes an interdisciplinary approach to teaching a linguistics and cryptology course. The authors, a Classics professor and a Mathematics professor, co-taught a three-week course, entitled "Deciphering Secrets Throughout History," to undergraduate students of varying backgrounds in mathematics and the humanities. Students were taught to apply tools from linguistics, statistics, and cryptanalysis to examine ancient texts, languages, and ciphers. The course culminated in an extended analysis of the fifteenth-century Voynich Manuscript, where students proposed their own original analyses of the text.

1 Introduction

In the past couple decades, a plethora of cryptology-related courses have emerged in the undergraduate curriculum. Many of these courses are traditionally taught in mathematics and computer science departments, with topics ranging from historical ciphers to post-quantum cryptography, and target audiences ranging from introductory students to graduating majors. Course topics, in-class activities, projects, and pedagogical approaches for teaching cryptology courses have been highlighted (though not exhaustively) in *PRIMUS* (Kaur, 2008), (Aydin, 2009), (Karls, 2009), (Schembari, 2020), *Cryptologia* (Winkel, 2008), (Kurt, 2010), (Glass, 2013), and previous HistoCrypt Conference Proceedings (Musílek and Hubálovský, 2018), (Krapp, 2019).

Some of these courses have gone beyond the typical mathematics and computer science audiences. For instance, Koss (2014) taught a first-year college writing seminar using cryptology as a vehicle for analyzing information literacy, critical thinking, and writing. In some cases, interdisciplinary courses co-taught by professors from two different disciplines have emerged. For example, Karst and Slegers (2019), an applied mathematics professor and a philosophy professor, co-taught a course partially focused on ethical issues in modern cryptography. Despite the wide range of interdisciplinary connections found within undergraduate cryptology courses, the literature lacks discussion on courses that have both a significant linguistics and cryptology focus. To help provide a novel contribution to the literature that fulfills this need, the authors, a Classics professor and a Mathematics professor, describe the structure of a three-week interdisciplinary linguistics and cryptology course that they co-taught and share some of their successful in-class activities.

2 Context

The setting for this unique interdisciplinary course was Furman University which is a small liberal arts university largely focused on undergraduate education. This school encourages interdisciplinary learning by annually offering a three-week May Experience (MayX) term after the spring semester ends. MayX classes must be non-traditional courses that are not taught during the semester, and they can be co-taught between two faculty members from different disciplines. While the authors have very different academic backgrounds (one is a theoretical mathematics Ph.D. whose research is in geometric topology but he has also taught some introductory college-level cryptology courses; the other is a Classics professor specializing in archaic Greek literature and historical linguistics, the study of language change, particularly within the Indo-European language family), co-teaching a MayX course that focused on applying tools from both cryptanalysis and linguistics in the context of ancient texts, languages, and ciphers seemed like a fantastic opportunity

to combine our skill sets and academic interests. This resulted in the creation of "Deciphering Secrets Throughout History," (referred to as Deciphering Secrets moving forward), which we cotaught in May 2019.

For such a course, we hoped to attract students from a variety of majors to encourage interdisciplinary learning. The twelve students that enrolled in the course had a wide range of background knowledge in mathematics, cryptology, and linguistics. As a result, we had to introduce basic background material in cryptology and linguistics, do our best to make immediate connections between these two disciplines, and leverage student skill sets to encourage productive group work.

3 Course Components

In this section, we will discuss the major topics in cryptology and linguistics covered in this course and how this material was synthesized into a cohesive structure. Topics were chosen based on the audience, time constraints, and our course goals, which included: acquaint students with the history, development, and methods of cryptology, and linguistics; apply tools from cryptanalysis and linguistics to analyze historical codes and texts in order to draw informed conclusions about the structure of these texts; further develop independent research skills and group work skills; reflect on and communicate the importance of interdisciplinary collaboration in academics and non-academic careers. Below, we describe the key introductory topics emphasized in each discipline separately before turning to how students were encouraged to synthesize these complementary fields in an activity involving the Voynich Manuscript.

3.1 Cryptology and Mathematics Components

For the cryptology components of this course, we focused on the Caesar shift cipher, monoalphabetic substitution ciphers (MSCs), nomenclators, and the Vigenère cipher. These topics were partially chosen because many of them introduced important cryptanalysis tools, such as frequency analysis, vowel recognition algorithms, and the Index of Coincidence, that could be applied more broadly to linguistic analysis. For most of these topics, our main resources were Bauer's undergraduate cryptology textbook *Secret History: the Story of Cryptology* (Bauer, 2013) and Singh's *The Code Book* (Singh, 2000).

The cryptology portion of our course started with the (Caesar) shift cipher, which is a basic substitution cipher where each letter in a text is replaced with another letter by shifting down the alphabet by a fixed amount. Because of the small key space (only 25 nontrivial keys when using the English alphabet), brute force attacks can easily be implemented to break shift ciphers. However, we challenged students to develop their own tactics for breaking shift ciphers by hand based on the linguistic structure of the underlying language and the rigid structure of shifting. To this end, we provided cryptanalysis activities where the plaintexts were written in a variety of languages (and without knowledge beforehand of which language was used): English, French, German, and Latin. This led to a discussion on differences in linguistic structures of these languages: most common letters, most common words and letters that begin a sentence, most common short words, etc., and how these differ among the languages.

We then transitioned to MSCs, which encompass any encryption method where each letter or symbol used in a text is replaced with one and only one letter or symbol. Unlike the shift cipher, the number of possible keys can be quite large for an MSC. For instance, if the English alphabet is used for writing, then there are 26! possible keys, making a brute force attack not possible. For cryptanlaysis, this naturally led to introducing frequency analysis and expanding upon the linguistic attacks used on shift ciphers. For this topic, we again wanted to give students examples of ciphertexts to break where they were required to analyze the linguistic structures of languages beyond English. One MSC example we used that met these goals was Challenge 1 of the 2016 Kryptos competition. This challenge featured an intercepted encrypted telegram between rumrunners in the 1920s where context can lead to conjecturing that the underlying language is Spanish, and context clues can assist with looking for key words in the text; see `http://www.cwu.edu/math/previous-challenges`, for the corresponding exercise and solution.

Beyond these traditional topics in MSCs, we also discussed nomenclators and vowel recognition algorithms. A nomenclator implements an MSC for parts of the encryption process, combined with a list of code words and symbols used

to replace more common words, bigrams, trigrams, and names, or represent nulls. This topic came up since we had students read parts of the *The Code Book* (Singh, 2000), and nomenclators are used in the Babbington Plot, which is discussed in Chapter 1. We also worked on implementing Sukhotin's vowel recognition algorithm, as described in Section 1.12 of Bauer (2013). Vowel recognition not only helps with breaking an MSC, but also can assist with conjecturing vowels in an ancient text where the underlying language or writing system might be unknown at the start. In particular, vowel recognition algorithms have been used to analyze the Voynich Manuscript, as noted in chapter 2 of Bauer (2017).

We next considered the Vigenère cipher, a polyalphabetic substitution cipher that requires a keyword to be shared between parties for encryption. This keyword designates a sequence of shifts to be used for substituting letters, with the keyword repeating as needed to complete the encryption process. For instance, if your plaintext said "Histocrypt" and your key was "dog," then a shift of $A \to D$ would be used to encrypt plaintext letter H to ciphertext letter K, a shift of $A \to O$ would be used to encrypt plaintext letter I to ciphertext letter W, and so on.

A major step in breaking the Vigenère cipher is finding the key length. We examined two methods for finding the key length: The Babbage–Kasiski test and the Index of Coincidence. Since the Babbage-Kasiski test did not come up in our Voynich Manuscript activity (see Section 3.3), we refer the reader unfamiliar with the technique to chapter 2 of (Singh, 2000). The index of coincidence (IC) measures the probability that two different letters chosen at random from a text will be the same. More formally, let N be the length of a given text and let F_α be the number of times a letter α occurs in that text. Then the index of coincidence for that text can be calculated as

$$IC = \sum_{\alpha \in \Omega} \frac{F_\alpha(F_\alpha - 1)}{N(N-1)},$$

where Ω represents the set of all letters in your alphabet, or more generally all graphemes in your writing system; see Section 3.2 for further discussion on linguistic terms. Since using a long keyword for the Vigenère cipher flattens the frequency distribution of the ciphertext letters, the IC can be used to provide an estimate on the key length.

In class, we covered the necessary background on combinatorics and probability to formally describe the IC and justify its connection to key length. We then worked on some basic examples of calculating the IC, where students created Excel worksheets to assist with calculations. Afterwards, students implemented both the IC and the Babbage–Kasiski test to assist with Vigenère cipher cryptanalysis exercises.

At the same time, by calculating the IC over numerous texts in a common language, one can calculate an expected value IC for that language. For instance, the expected value IC for English is approximately 0.0667, while the expected value IC for Spanish is approximately 0.0775; see Chapter 2 of Bauer (2013) for some more examples. Thus, this was a tool that naturally fit into our course since it could be applied to both breaking the Vigenère cipher (and more broadly, any polyalphabetic substitution cipher) and analyzing ancient texts to make conjectures about what language underlies a text or if the text was possibly encrypted using certain methods.

Our final mathematical topic was entropy, which supplies a statistical measure of information contained in a text or language. This concept was first introduced by Shannon (1948) and we refer the reader to Section 11.2 of Bauer (2013) for an introduction to this topic. One can theoretically compute the (expected value) entropy of a language as

$$H = -\sum_{i=1}^{n} M_i \log_2(M_i),$$

where M_i indicates the probability of message i, and the summation is taken over all possible messages in that language. In practice, such a calculation is unreasonable to compute. However, it is reasonable to calculate the n^{th}-order entropy, H_n, for n sufficiently small, which provides an estimate for H. For instance the first-order entropy of English can be calculated as $H_1 = -\sum_{i=1}^{26} p_i \log_2(p_i)$, where p_i is the probability of the i^{th} letter of English occurring on average in a text written in English. Similarly, H_2 is an expected value where probabilities of bigrams in the relevant language are used, and so on. In a similar manner, one can calculate the n^{th}-order entropies of a given text and make a comparative analysis with the expected value entropies of languages and other texts. See Bennett (1976) for a chart of first-,

second-, and third-order entropy values for various languages and writers.

While we did not apply ideas from entropy to any cryptanalysis assignments, this topic naturally built off the probability background established from defining the IC and the use of frequency analysis in breaking MSCs. The only mathematical background needed to be introduced (or reviewed) were properties of logarithms. Like the IC, n^{th} order entropies can be used to conjecture if a text has been encrypted and which languages possibly underlie a text. Furthermore, entropy has been applied as a powerful quantitative tool in linguistic analysis, including the Voynich Manuscript, making this an ideal topic for our course.

3.2 Linguistics and History Components

To complement the cryptological principles, tools, and historical ciphers emphasized in the previous section, the linguistics component of the class introduced core concepts from the basic subfields of phonology (study of sounds), morphology (study of word form), and grammatology (study of writing systems and scripts). We also highlighted historical texts and examples of scripts and their decipherment through linguistic analysis, particularly the cases of Egyptian hieroglyphics and Linear B, as well as scripts still yet to be deciphered, including Linear A. Before discussing how these famous examples were deciphered or have been attempted to be deciphered, students were first introduced to the basic building blocks of language and writing.

We began with the question of what makes a language a language, and how spoken languages in particular are based on a system of sounds, as opposed to sign languages which are based on a system of gestures. Spoken languages rely on the human vocal tract's ability to produce sounds, which can be divided into two essential categories, vowels and stops (consonants). When combined, consonants and vowels form sound sequences in human speech that we call syllables. Syllables in turn serve as the phonological building blocks of words. For the purposes of time, we did not go into detail about additional types of sound distinctions that some languages employ, such as stress and pitch, among others. It was important to establish a basic understanding of vowels, consonants, and syllables before discussing how writing works. Starting in this way also reinforced the class's grasp of vowel recognition algorithms aiding in decryption.

Our next unit turned to the issue of writing systems and how they encode the sounds and ideas of a language through graphemes (or characters), which are the most basic contrastive unit of a writing system. We then provided a historical survey of the world's writing systems and the basic typology used by linguists to categorize them; see Daniels (1990) and Rogers (2005) for a more detailed overview of writing systems. All writing systems can be categorized as one of the following:

1. morphosyllabary: each grapheme stands for the sound of a morpheme, the most basic meaningful unit of a language (e.g. Chinese characters, or hànzì, for Chinese)

2. syllabary: each grapheme stands for a syllable (e.g. Katakana for Japanese)

3. abjad: each grapheme stands for a consonant only, while no vowels are represented (e.g. Hebrew and Arabic, although these are no longer pure abjads)

4. abugida: each grapheme stands for a consonant, while additional flourishes may be added to the character to indicate particular vowels (e.g. Devanāgarī for Sanskrit)

5. alphabet: each grapheme stands for either a vowel or a consonant (e.g. Latin alphabet for English, French, and many more)

6. featural script: each grapheme represents a phonological feature of a sound segment (e.g. Hangul for Korean)

As these writing systems represent different segments of sounds, each system consists of a different number of graphemes. Syllabaries, for instance, tend to feature a high number of graphemes, ranging from 50 to 200 characters, while alphabets tend to feature a lower number, ranging from 20 to 40 characters. Knowledge of these statistics helped the students to predict what type of writing system they were encountering, even if they had never been exposed to the script before. We also considered how each writing system posed different challenges and advantages to codebreaking methods based off frequency analysis. For instance, students needed to make the connection that the larger the set of graphemes in

a writing system, the longer a text would need to be in order to approximate the average frequencies of the underlying language.

With both the types of world writing systems and the types of graphemes employed for each system established, we turned to other key elements of writing, including orientation of writing, syllable division, and word division. We discussed what visual cues we might use to determine each of these elements. We tasked students with analyzing known scripts (although the students did not necessarily recognize all of them) and identifying their graphemes, writing orientation, syllable/word divisions, and ultimately writing system type. The students had to explain concretely in linguistic terms how they were able to discern each of these elements. Examples of scripts we used for this exercise are provided in Figure 1 below.

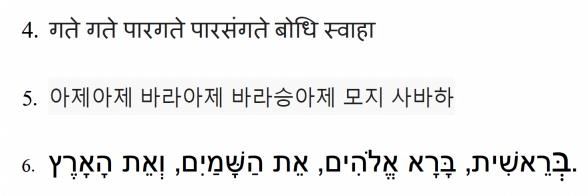

Figure 1: Writing Systems Exercise

In item 4 of the figure above, for example, students were expected to recognize distinct recurring graphemes, to which some alterations were made above a key horizontal line, which should allow students to infer that they were dealing with an abugida script type.

We also considered why different languages adopted a particular writing system. This served as an introduction to the important concept of language classification by language families, which refer to groups of languages deriving from a common ancestral language. Some of the most well-known and well-established language families include Indo-European, Sino-Tibetan, and Afro-Asiatic, among many more. Semitic languages such as Hebrew and Arabic, a subset of the Afro-Asiatic family, have tended to employ abjad scripts, which lend themselves well to encoding the Semitic word structure, which is heavily based on a triconsonantal root and vowel patterns. Indo-European languages such as Greek and Latin, on the other hand, have less predictable vocalizations and have therefore generally eschewed abjad scripts in favor of other segmental scripts like the alphabet, since they can more clearly represent and distinguish vowels and consonants in writing. With an understanding of which writings systems different languages tend to employ, students could now make good hypotheses about potential underlying languages of an unknown script based purely off the type of writing system being used. For instance, if students identify an unknown script as an abjad type, they might reasonably assume the underlying language to be Semitic.

At this stage, students had a good grasp of essential linguistic concepts to consider codebreaking in a new light, and thus we prepared them to begin synthesizing the cryptological and linguistic components of the class by showcasing different historical examples of script decipherment. Our first case study was the decipherment of Egyptian hieroglyphics, which was an example where researchers did not recognize the script but knew the underlying language (Egyptian). The successful decipherment of hieroglyphics was ultimately made possible by the existence of the Rosetta Stone, which also included Demotic and Ancient Greek inscriptions that presumably translated the hieroglyphics, and hence provided a crib for the unknown script; see Robinson (2012) for a detailed history of this decipherment. The French historian and linguist Jean-François Champollion ultimately cracked the writing system by understanding that what appeared to be pictographs (drawings representing ideas) actually represented sounds. He was able to match the glyphs to particular sounds by linking cartouches to the phonetic representation of names such as Alexander (Alexandros) and Ptolemy (Ptolemaios), which appeared in the Demotic and Ancient Greek inscriptions on the Rosetta Stone. Names being names cannot be translated but must rather be phonetically mimicked in other languages. This historical example showcases the problems that arise in considering what a particular grapheme might represent: a vowel, a consonant, a syllable, or even an entire word or idea, in which case it might not be possible to discern the sounds of a language that correspond to a word represented in a pictograph.

Students were then tasked with identifying sound values for Egyptian hieroglyphics, which is primarily an abjad that does not represent any vowels in its glyphs, through a worksheet where we included cartouches recording the names Ptolemy, Berenike, Cleopatra, and Alexander.

This exercise provided a language complement to MSC problems emphasized during the cryptology-focused portion of the course. In both cases, students had to determine a direct one-to-one correspondence between two separate value sets, whether it was matching a sound to a grapheme or matching plaintext letters to ciphertext letters. The exercise also built upon previous linguistic concepts, since students once again had to isolate what constituted a single grapheme and figure out how these graphemes combined together to produce words. We provide below in Figure 2 the cartouches for Cleopatra (left) and Alexander (right) used for the worksheet.

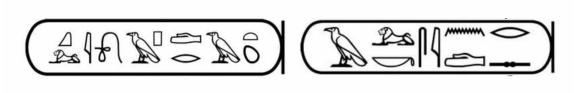

Figure 2: Egyptian Hieroglyphics Exercise

Our second case study in decipherment was that of Linear B, an ancient syllabic script that was preserved on baked clay tablets and used to record Mycenaean Greek in the Bronze Age. This example is significant in that researchers did not know the script or the underlying language recorded by it, nor were there any known translations that could serve as a potential crib. The decipherment of Linear B thus offers one of the most exciting examples of linguistic breakthroughs, which was made possible through the combined efforts of three individuals (Alice Kober, John Chadwick, and Michael Ventris), who each put forth unique contributions based on their professional experiences as a classicist, codebreaker and linguist, and architect; see Chadwick (1958) for a thorough overview of the Linear B decipherment. In particular, statistical methods (essentially frequency analysis) revealed underlying patterns of regularities that made it possible to associate phonetic and semantic values with the symbols. Through a form of brute-force attack, syllabic sound values were variously cross-checked (purely by guessing at first) across all the symbols until the combination of sounds for a particular word produced recognizable place names in Greece. Once the sound values for just a few symbols of the Linear B syllabary were established, the remaining number (with a few exceptions) were soon after resolved. Even before the sound values could be determined, linguistic analysis of the script had already proved that the encoded language was inflected. Linear B words with the same stable set of initial symbols (i.e. word stems), but with regular changes to their endings, signalled changes in word form that appeared to indicate different grammatical functions (i.e. inflection). This breakthrough allowed linguists to reasonably assume that the underlying language was Indo-European, and through knowledge of the historical development of sounds from Proto-Indo-European, the underlying language was ultimately determined to be Mycenaean Greek. The incorporation of both statistical and linguistic considerations was essential for developing a thorough understanding of the language and script of Linear B. This historical example modelled the type of interdisciplinary work we encouraged our students to emulate in our class.

To simulate the processes involved in the decipherment of Linear B, we created an exercise that involved deciphering a fictional script called Linear C. This script encoded another fictional language called Yomama. Students were provided a list of symbols that could be determined to be words through writing, but their pronunciation and meaning were still unknown. The words are provided in Figure 3 below.

a) ♥ ◊ ⊗ ♣

b) ♥ ◊ ♠ ◊

c) ♥ ◊ ♣ ∇

d) ∅ ♣ ♥ ◊

e) ∅ ♣ ⊕ ∇

f) ∅ ♣ ∇ ♣

g) ♦ ♦ ◊ ∇

h) ♦ ♦ ∅ ♣

i) ♦ ♦ ♦ ◊

Figure 3: Linear C Data

Students were then provided with an additional piece of data: a recently unearthed but fragmentary text revealing that the Yomama word for gold was pronounced *gobade*. We also shared the word in Linear C, provided in Figure 4 below. The students were tasked with finding the phonetic values for all the symbols of Linear C and were told to

⊗ ♥ ◊

Figure 4: Linear C Problem

assume the following: There are three consonants and three vowels, giving Yomama the simplest phonological system known; reading and writing orientation is left-to-right; every symbol stands for a CV (consonant-vowel) syllable; if two distinct symbols share a consonant, they must differ in vowels; if two distinct symbols share a vowel, they must differ in consonants; all words consist of a stem and suffix; stems are of the form CVCVC; all suffixes are of the form VCV, and it may be assumed that suffixes sharing their final syllable are of the same suffix.

The goal of this exercise was to get students to recognize the unique challenges posed by a syllabic script, since the individual graphemes do not easily reveal the vowel sounds that may be used for each syllable. But by a systematic approach modelled by the example of the decipherment of Linear B, the full CV sound values for each symbol in the Linear B syllabary could still be identified. Students were then able to better understand both the linguistic and cryptological implications of different writing systems that encode sounds in different ways, specifically how consonants and vowels may be variously encoded in writing (or not), and how this may ease or frustrate attempts to decipher an unknown script or to decrypt a challenging code.

In conclusion to the linguistic portion of the class, we shared some additional case studies of scripts that still have yet to be deciphered, such as Linear A. In this case, the script is the same as Linear B; however, the underlying language is unknown. All attempts to decipher it have been unsuccessful so far; see Salgarella (2020) for an overview of the unique challenges posed by Linear A. Ultimately, for such cases, we don't have enough surviving textual examples or data to allow for a securable decipherment. Exposure to this example encouraged students to consider the minimum amount of text necessary to allow for successful decipherment or decryption.

The goal of studying all of these historical case studies was to demonstrate that methods employed to decipher an unreadable script can also be employed to break a secret code or cipher and vice versa. Students could now attempt to approach a practical and famously longstanding problem of decipherment/decryption themselves.

3.3 Course Synthesis: The Voynich Manuscript Activity

Following the instruction of the individual components of cryptology and mathematics on the one hand, and linguistics and history on the other, the authors devised an original group activity that synthesized all of these topics as the culminating course experience. This activity required students to conduct an original analysis of the Voynich Manuscript (hereafter referred to as VM), a fifteenth-century illustrated codex hand-written in an unknown writing system with an equally unknown language underlying the script. A printed color copy of this manuscript consisting of all 116 folios (leafs, or pages, of the manuscript) can be found in Clemens (2016) and we refer the reader to Figure 5 for a visual of a folio. Though the VM has defied all attempts at decryption, it provided the perfect practical testing ground for the students to experience firsthand an attempt at codebreaking and script decipherment through interdisciplinary collaboration and synthesis of their newly acquired skill sets.

For this assignment, students worked in groups of three to perform a statistical and linguistic analysis of a folio of the VM, applying the different methods and concepts they had learned throughout the course. At the end of the analysis stage, each group gave a ten-minute presentation to share their conclusions. Four folios were preselected for the students to examine: folio 42 recto, folio 81 recto, folio 93 recto, and folio 99 verso. These four were selected due to the clearly discernible writing components and additional remarkable elements of each folio, particularly illustrations, which could provide useful context to the text and further aid the students in their original analyses. Students were able to examine these four folios in-depth through the high-definition images provided in `https://www.jasondavies.com/voynich/#f1r/0.5/0.5/2.50`. We also shared additional online resources to aid in their analysis (see Section 3.4 below).

The students were first tasked with isolating the set of graphemes that appeared in their folio. They had to consider how many unique graphemes they could identify, and what type of writing system

Figure 5: Folio 42 recto from the Voynich Manuscript

the number of existing graphemes might indicate. They also took care to note any patterns in the use and appearance of a single grapheme in the text, particularly if it appeared in a restricted environment, such as in word-initial or word-final positions.

Historically, this grapheme identification process for the VM has been a challenging but important task, since so many of the statistical tools discussed in Section 3.1 are dependent on having a clearly defined set of graphemes. There are theories and evidence that the manuscript consists of different sections authored by different individuals; see Chapter 2 of Bauer (2017) for more details. This challenge led to a productive class discussion on how many of the cryptanylsis tools we had discussed (e.g.frequency analysis, vowel recognition algorithms, index of coincidence, and entropy) could be applied to help identify if a text was encrypted, the type of encryption, or the underlying language. But these tests are all dependent on first identifying your set of graphemes. Thus, this led to a clear order of operations for an analysis of the VM: students first needed to identify the graphemes for the folio under analysis, and then look to apply the relevant statistical tools from cryptanalysis to draw conclusions about the text.

After this initial linguistic analysis, the students provided arguments and counterarguments for the use of a particular writing system being employed in the VM, considering first if it was a phonetic writing system and if so, which one: abjad, abugida, alphabet, or featural.

The next stage of analysis involved calculating the grapheme frequencies of their assigned folio. The students first created their own transliteration systems to make the VM text machine-readable and recorded their graphemes in an Excel sheet. Using their newly created datasets, they made the appropriate calculations, including the first order entropy and the index of coincidence for their folio. Results were interpreted through comparison to the entropies and indices of coincidence calculated for other known languages that they thought might be the underlying language of the VM. The students were also asked to consider what statistical tests other than average word length and word length distributions they might apply to their folios, and what complications might arise.

Following this close study of a single folio, the students expanded their investigation to the rest of the VM and attempted to find another folio that might come from a different hand, use a different writing system, or record a different language. Students had to justify their reasoning if they did or did not find evidence for different languages and writing systems within the VM. Finally, the students drew conclusions on what they thought the text of the VM represented at this point: a hoax, an MSC text, a PSC text, a universal language, a natural language (and if so, which language family), multiple languages, or some combination of these options.

This activity was clearly the highlight of the course, with students noting that this was a concrete area where they could enjoy and experience firsthand the rewarding coordination and synthesis of cryptology and linguistics.

3.4 Voynich Manuscript Worksheet

Below, we include the setup and questions from our VM activity (without commentary) to assist educators interested in incorporating or modifying this activity.

Directions: For this assignment, each student will work in a group of three to perform a statistical and linguistic analysis on a folio of the Voynich Manuscript. At the end of this analysis, each group will have 10 minutes to share their work with the class. Use

the link https://www.jasondavies.com/voynich/#f1r/0.5/0.5/2.50 to find the necessary folios of the manuscript. Group 1 is assigned Quire 6 f42r, group 2 is assigned Quire 13 f81r, group 3 is assigned Quire 13 f93r, and group 4 is assigned Quire 19 f99v.

You may find the following links helpful for your analysis:

https://voyant-tools.org/ - text analysis, including word frequencies

http://textalyser.net/ - letter frequencies, word frequencies, word length (don't rely on syllable count)

https://md5decrypt.net/en/Letters-frequency-analysis/ - frequency count by letters, digraphs, trigraphs, etc., with comparisons to other languages

1. For homework, each member of your group should have determined the set of graphemes for your folio.

   a. Decide on a single set of graphemes for your group that you will use for the rest of your analysis. How many unique graphemes did your group identify for your folio of the VM?

   b. Which kind of writing system does the number of existing graphemes indicate?

   c. Provide arguments and counterarguments for the use of the following writing systems in the VM:

   i. Abjad

   ii. Abugida

   iii. Alphabet

   iv. Featural

2. Calculate the grapheme frequencies of your assigned folio. You can use Excel or other programs listed above. Also, it might help to develop a transliterative system and make the VM text machine-readable.

3. Calculate the first order entropy (H1) and the index of coincidence for this text and interpret your results. Compare these calculations to those of known languages that are possibly the underlying language of the VM.

4. What other statistical tests could you run to gather data on your page? For instance, your first reading on the VM showed comparisons on average word length and word length distributions to draw some conclusions about the script. Come up with one other statistical test that could be useful and apply that to your folio. Interpret your results and draw comparisons to other languages.

5. Do you notice any patterns in the use and appearance of a particular grapheme in your text? Do any graphemes appear in a particularly restricted environment, such as word-initial or word-final positions only? What can you conclude from the observation of such occurrences?

6. Look through the manuscript and find one other folio that your group thinks was written by the same hand, using the same writing system, and the same language. Justify your reasoning.

7. Look through the manuscript and find one other page that you think either came from a different hand, a different writing system, or uses a different language. Justify your reasoning. If no such folio exists, explain why.

8. If you wanted to do a more thorough statistical analysis of this text, what would you do? What complications might arise?

9. Based on your analysis of the VM with your team, what do you think the text represents at this point: a hoax, an MSC text, a PSC text, a universal language, a natural language (and if so, which language family), multiple languages, or some combination of these (specify)?

4 Conclusion

This paper has demonstrated how a cryptology course taught through multiple disciplinary perspectives can contribute to the current range of pedagogical approaches employed at an undergraduate institution. An interdisciplinary approach holds great appeal for students of broad disciplinary backgrounds and interests, and offers a promising way to enrich current undergraduate course offerings that focus exclusively on cryptology or linguistics in separate courses. The thoughtful implementation and combination of different disciplinary skill sets to the same problem can enrich student engagement, facilitate col-

laborative learning, and raise greater metacognitive awareness of undergraduate learning and its applicability to practical problems.

Finally, while this course did have a specialized format, there are elements that could easily be transferred to a variety of introductory college classes. Since no prerequisites were required for our students and there was not a significant amount of content covered from any one discipline, our major course activity on the Voynich Manuscript (discussed in Section 3.3) could easily fit into an introductory cryptology course taught in a mathematics or computer science department, a linguistics course, or as a student project in any such courses. We hope this paper inspires other educators to incorporate interdisciplinary approaches into their cryptology and linguistics courses.

Acknowledgments

This work was funded by Furman University's Interdisciplinary May Experience Course initiative. The authors would like to give special thanks to Christopher Hutton and the MayX Committee for their guidance and support.

References

Nuh Aydin. 2009. Enhancing undergraduate mathematics curriculum via coding theory and cryptography. *PRIMUS*, 19(3):296–309.

Craig P. Bauer. 2013. *Secret history*. Discrete Mathematics and its Applications (Boca Raton). CRC Press, Boca Raton, FL. The story of cryptology.

Craig P. Bauer. 2017. *Unsolved! The history and mystery of the world's greatest ciphers from ancient Egypt to online secret societies*. Princeton University Press, Princeton, NJ.

William Ralph Bennett, Jr. 1976. *Scienfitif and Engineering Problem-solving with the Computer*. Prentice Hall, Englewood Cliffs, NJ.

John Chadwick. 1958. *The Decipherment of Linear B*. Cambridge University Press, Cambridge, UK.

Raymond Clemens, editor. 2016. *The Voynich Manuscript*. Yale University Press, Yale, CT.

Peter T. Daniels. 1990. Fundamentals of grammatology. *Journal of the American Oriental Society*, 110(4):727–731.

Darren Glass. 2013. A first-year seminar on cryptography. *Cryptologia*, 37(4):305–310.

Michael A. Karls. 2009. Codes, ciphers, and cryptography—an honors colloquium. *PRIMUS*, 20(1):21–38.

Nathaniel Karst and Rosa Slegers. 2019. Cryptography in context: Co-teaching ethics and mathematics. *PRIMUS*, 29(9):1039–1059.

Manmohan Kaur. 2008. Cryptography as a pedagogical tool. *PRIMUS*, 18(2):198–206.

Lorelei Koss. 2014. Writing and information literacy in a cryptology first-year seminar. *Cryptologia*, 38(3):223–231.

Peter Krapp. 2019. Beyond schlock on screen: Teaching the history of cryptology through media representations of secret communications. *Proceedings of the 2nd Conference on Historical Cryptology, HistoCrypt 2019*, pages 79–85.

Yesem Kurt. 2010. Deciphering an undergraduate cryptology course. *Cryptologia*, 34(2):155–162.

Michal Musílek and Stepán Hubálovský. 2018. Teaching and promoting cryptology at faculty of science university of hradec králové. *Proceedings of the 1st Conference on Historical Cryptology, HistoCrypt 2018*, pages 137–143.

Andrew Robinson. 2012. *Cracking the Egyptian Code: The Revolutionary Life of Jean-Francois Champollion*. Oxford University Press, Oxford, UK.

Craig P. Rogers. 2005. *Writing Systems: A Linguistic Approach*. Blackwell Publishing, Malden, MA.

Ester Salgarella. 2020. *Aegean Linear Script(s): Rethinking the Relationship Between Linear A and Linear B*. Cambridge University Press, Cambridge, UK.

N. Paul Schembari. 2020. A half-rotor cipher for the classroom. *PRIMUS*, 30(5):552–570.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656.

Simon Singh. 2000. *The Code Book: The Science of Secrecy from Ancient Egypt To Quantum Cryptography*. Anchor, New York, NY.

Brian Winkel. 2008. Lessons learned from a mathematical cryptology course. *Cryptologia*, 32(1):45–55.