What is the Code for the Code? Historical Cryptology Terminology

Vasily Mikhalev[1], Nils Kopal[1], Bernhard Esslinger[1], Michelle Waldispühl[2], Benedek Láng[3], Beáta Megyesi[4]
[1]University of Siegen, Germany
[2]University of Gothenburg, Sweden
[3]ELTE, Hungary
[4]Uppsala University, Sweden

Abstract

The cross-disciplinary nature of historical cryptology involves the challenge to find a terminology that is both consistent and accepted across the different disciplines and applicable in the single fields. In this paper, we propose a terminology based on concise principles developed by an interdisciplinary group of researchers. We present terms prominent in the study of historical cryptology, define them, and illustrate their usage. Our goal is to initiate and/or continue the discussion of how we use various terms for different types of historical encrypted sources, and their study. Our hope is that this paper will contribute to consistent and systematic usage of terms in the HistoCrypt community.

1 Introduction

Historical cryptology, the study of codemaking and codebreaking of historical ciphers, is a cross-disciplinary field engaging not only cryptographers and cryptanalysts but historians, linguists, computational linguists, computer scientists, computer vision specialists, codicologists, paleographers, archivists, and librarians, to name a few. Each field has its own angle and methodology to find the answers to the research questions of their interest. This might include the study and the interpretation of the ciphers, ciphertexts, codes, keys, nomenclators, nomenclatures, or codebooks. The task is not easy given the wide range of time periods, geographic areas and languages covered by the sources.

The encrypted material evolved over the centuries; many types of linguistic entities have been encrypted from letters, syllables and morphemes to named entities, words and phrases, with different code structures including various alphabets, digits, or graphic signs with fixed and/or variable length of codes.

Over the years, we have seen numerous studies dealing with historical encrypted sources, many with their own usage of specific terminology, defined or left to be interpreted by the readers. The problem is further complicated by the fact that the meaning of terms has changed over time and some terms have multiple meanings not only across but also within the same study. Additionally, variation in British and American English might also create confusion.

In light of the above mentioned reasons and challenges, we present hereby a proposal of terms and their definitions for describing the most common concepts related to ciphertexts on one hand, and cipher keys on the other. Our long-term goal is to create a consistent terminology for historical cryptology which fits various scientific fields involved and which covers and allows for expressing the most common concepts in our field.

While there have been previous attempts to introduce more or less consistent terminology for historical cryptology, such as (Meister, 1906; Friedman, 1959; Employees of Bletchley Park, 1945; Kahn, 1996; Schmeh, 2018; Dunin and Schmeh, 2020), we believe our proposal is unique in its actuality, and its well-defined structure grown out to be a compromise between experts from various scientific disciplines. Our aim was not to reconstruct the historical actors' categories, i.e. how they referred to the various elements of the encryption process. Nor was it to create a terminology that is primarily applicable to modern ciphers. Rather, we aimed at introducing consistent and modern terminology that is applicable to the historical ciphers. However, our aim is not only to make historical-cryptology terminology consistent, but also adequate, unambiguous, and simple to use in order to be able to become a standard. To achieve our goals, we tried to be

specific without being too complex so that people without a background in the field can read and hopefully also apply the terminology suggested in our work.

Last but not least, we would like to encourage the community to continue the discussion about terminology issues. Our hope is that the community will adapt the proposed terms systematically in the future whenever suitable and appropriate. Needless to mention, we are open to changes and welcome feedback and suggestions for improvements. After all, standards are not given from scratch but emerge by systematic usage by many people.

In the following, we start by presenting previous attempts to describe terminology related to historical cryptology. In Section 3, we describe the principles behind our proposal, followed by a description of the usage of the terminology. In Section 4, we introduce the terms with their definition, and in Section 5 we discuss our reasoning, problems and some shortcomings of our approach. Lastly, in Section 6, we conclude the paper and give some directions for future work.

2 Related work

In this section, we present related work in the field of terminology for historical cryptology. Various researchers, authors, and cryptanalysts faced the same problem we did. They were writing about historical cryptographic topics or are part of cryptologic history themselves (e.g. because they worked in Bletchley Park). A metalanguage was needed for the description and study of historical documents, and many developed their own terminology or even applied terms without explicitly defining them. In this section, we briefly present the most prominent and important examples to the best of our knowledge.

In 1906, Aloys Meister wrote the most comprehensive collection and analysis of Papal ciphers in his German work "Die Geheimschrift im Dienste der päpstlichen Kurie von ihren Anfängen bis zum Ende des XVI. Jahrhunderts" (Engl. "The secret writing in the service of the papal curia from its beginnings to the end of the XVI century") (Meister, 1906). Mainly focusing on papal ciphers, he used terms like "Geheimschrift" (Engl. "secret writing" or simply cipher), "Nomenklator" (Engl. nomenclator), and "Trugbuchstaben" (Engl. letters of deception) which we today know as "nulls".

The "Bletchley Park Cryptographic Dictionary" (Employees of Bletchley Park, 1945) from 1944 is another work that introduces terminology of cryptology. A text reproduction of this dictionary can be found on Tony Sale's webpage[1]. The dictionary features a broad set of terms used by Bletchley Park employees in their daily work, e.g. Bombe (a cryptanalyical machine to break daily keys used with the German Enigma) or Tunny (a "German electric letter-subtractor, or virtual letter-subtractor, cipher machine using the teleprinter alphabet"). As can be seen, many terms were developed and listed quite specifically for analyzing World War II cipher machines.

William Friedman was one of the first who considered cryptology as a scientific field in its own right. He developed the idea that cryptology consists of two (main) parts: cryptography, which is the making of ciphers, and cryptanalysis which is the breaking of ciphers. Furthermore, he defined other important terms as part of cryptology, such as traffic analysis which is the analysis of communication flows. He wrote two book series: "Military Cryptanalysis" (Friedman, 1959) as single author, and "Military Cryptanalytics" (Friedman and Callimahos, 1985) which he co-authored with Lambros D. Callimahos. The books of both series were classified and only published for government use in the past. The "Military Cryptanalysis" series as well as the first two books of the "Military Cryptanalytics" series have been declassified. In "Military Cryptanalytics", a comprehensive glossary of the terms used in cryptology is presented. Their terminology was created for training of NSA and military cryptanalysts for the cryptanalysis of military ciphers.

David Kahn's "The Codebreakers" (Kahn, 1996) first published in 1967 is one of the most famous standard works on historical cryptology. It has inspired many researchers to become passionate about historical cryptology. In the introductory chapter of his book, Kahn introduces his terminology. He states that "Cryptology is the science that embraces cryptography and cryptanalysis, but the term "cryptology" sometimes loosely designates the entire dual field of both rendering signals secure and extracting information from them" (Kahn, 1996). Kahn introduces many, nowadays standard, and widely used cryptologic terms, e.g.

[1]The 1944 Bletchley Park Cryptographic Dictionary: https://www.codesandciphers.org.uk/documents/cryptdict/

plaintext, ciphertext, and cipher. Moreover, he introduces different types of ciphers such as monoalphabetic and polyalphabetic ciphers. His terminology is, of course, mainly needed to describe the historical cryptologic methods and practices presented in his book.

Two authors who started to use mathematical terms describing classical ciphers are Alan G Konheim (Konheim, 1981) and F L Bauer (Bauer, 1997).

Recently, Klaus Schmeh presented an overview of relevant terms and definitions in his blog (Schmeh, 2018) and later in the glossary of his and Elonka Dunin's book *Codebreaking: A Practical Guide* (Dunin and Schmeh, 2020). Schmeh was one of the first to point out the lack of consistency in the terminology of historical cryptology and made significant contributions to raise awareness for terminological issues and to initiate a discussion of standardization of terminology in the field.

Another glossary for Historical cryptology is available on the "Portal of Historical Ciphers" (Antal, 2018), which is a website developed and maintained by Eugen Antal, where a database of historical ciphers and keys, as well as tools for document analysis are provided. The terms presented in the glossary (Antal and Zajac, 2020) are taken from the aforementioned works by Schmeh (Schmeh, 2018) and Friedman (Friedman and Callimahos, 1985).

The release of the DECODE database (Megyesi et al., 2019) including a large collection of historical ciphers and keys with a description of the metadata about their origin, source, and characteristics led to the introduction of new terms, and the refinement of some others. For example, the distinction between plaintext (the underlying non-encrypted text) and cleartext (a non-encrypted part in the encrypted document) was suggested which is established today.

The transcription guidelines developed for ciphertexts and keys (Megyesi, 2020) and (Megyesi and Tudor, 2021) provide a further attempt to explain important terms of the field, but from a visual and paleographic (i.e. the study of handwriting) point of view to suggest consistent transcription of images of encrypted sources.

Important terms used in practice for explaining the structure of keys and the cryptanalysis of ciphertexts in the Papal correspondence during the 16th and 18th centuries in the Vatican have been introduced and explained in more or less detail in the paper by (Lasry et al., 2020).

Another recent article dealing with the study of the evolution of cipher keys presented an extensive description of the content of cipher keys originating from early modern times. The paper describes the plaintext as well as the code structure providing detailed descriptions of the content of keys (Megyesi et al., 2022).

Chapter 2 of the book (Esslinger, 2023), written by the same authors as this paper, describes historical cryptology and discusses the corresponding terminology.

The work and the terminology presented above serve as the basis for our suggestion for terminology for historical cryptology, which we describe next.

3 Creating terminology

Introducing and defining terms to create a nomenclature or terminology for scientific fields requires expert knowledge. Identifying frequently used terms in various contexts and interpretations as well as knowing the uncommon terms are indispensable. In order to succeed in acceptance by the public, adapting the terms to readers of various backgrounds and scientific fields is as important. Below, we reveal our reasoning and considerations that finally led to the principles we applied when developing the terminology for historical cryptology.

3.1 Principles

The main documents in consideration of our study are the plaintext, the ciphertext, and the cipher key. When we describe the work related to encrypted sources, we think of two sides of the coin: the ciphertext side with the code structure, and the plaintext side with the underlying text message consisting of linguistic entities. Our suggestion for terminology and the following five basic principles behind are based on the two different parts.

Symmetry Historical cryptology primarily focuses on ciphers, which are algorithms that convert plaintext to ciphertext by applying encryption and back to plaintext through decryption. One can easily see that this process is somewhat symmetric, which means that most of the concepts discussed in the scope of historical cryptology usually have their related expression on the other side

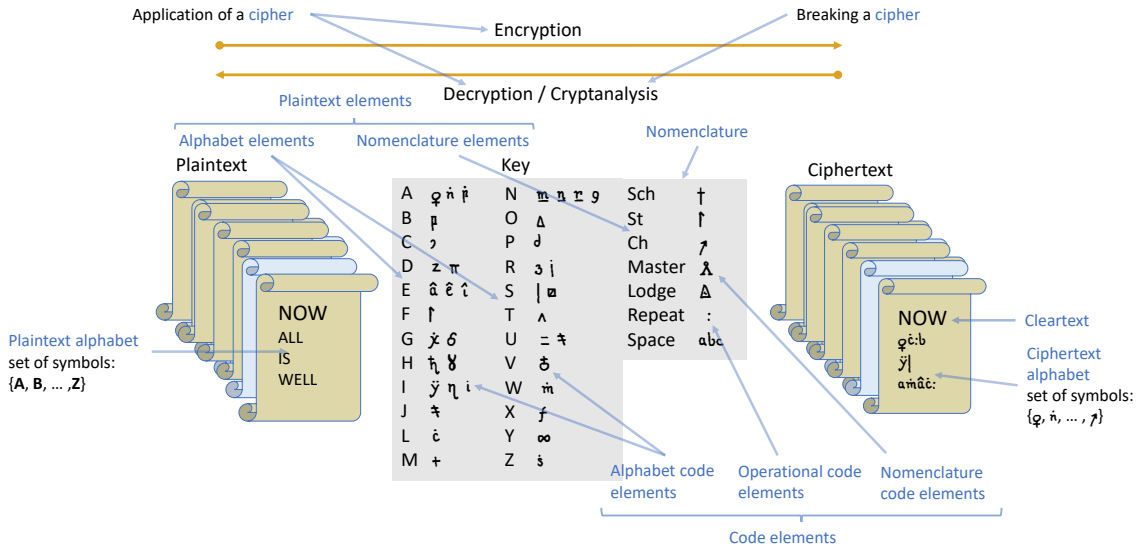

Figure 1: Most of the terms indicated together

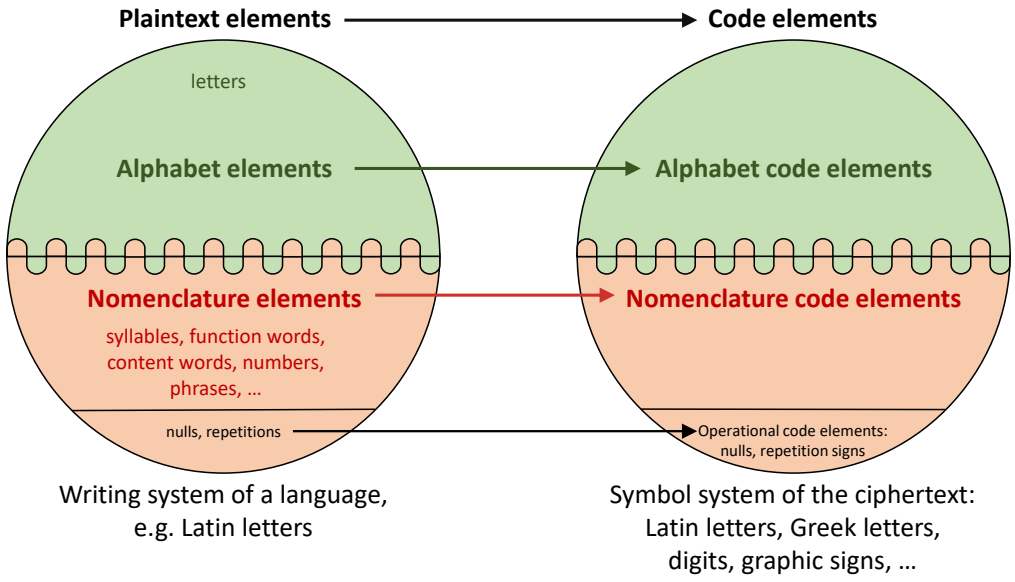

Figure 2: Mapping of the corresponding terms

of the encryption/decryption process. The goal is to ensure that the proposed terms adhere to this principle, which helps to create a well-defined structure for the terminology, as illustrated in Figures 1 and 2. In other words, we tried to think of terms in terms of pairs: on the ciphertext side, and on the plaintext side.

Explicitness Despite making sure that the related terms for each of the sides are proposed, it's also helpful if the one who sees the term for the first time immediately has a clear understanding to which side it refers. We use the term "code" specifically when referring to elements that are present in the ciphertext side, in order to provide immediate clarity and understanding for readers.

Hierarchy While developing the proposed terminology, we aimed to create a more organized system by clearly indicating when a group of elements is a subgroup of a larger group. Whenever possible we try to show the relationship between the terms of the same side, as represented in Figures 1 and 2.

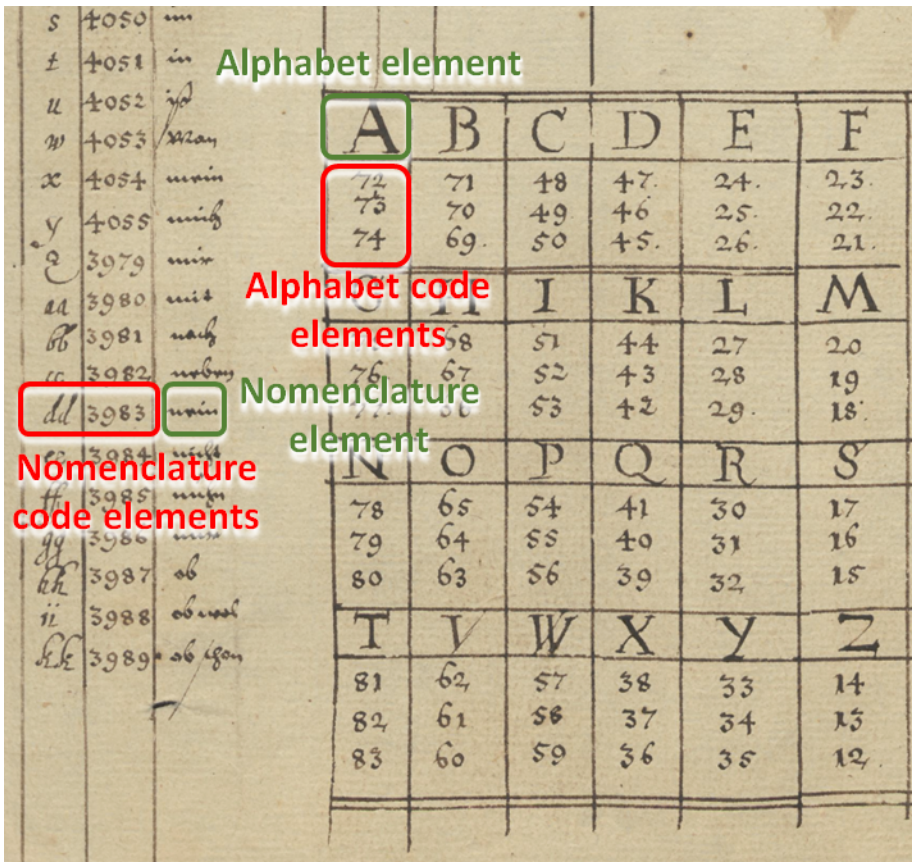

Figure 3: Example of a key: key used in Swedish diplomatic correspondence in the 1630ies, Riksarkivet Sweden, Chifferklaver II:24.

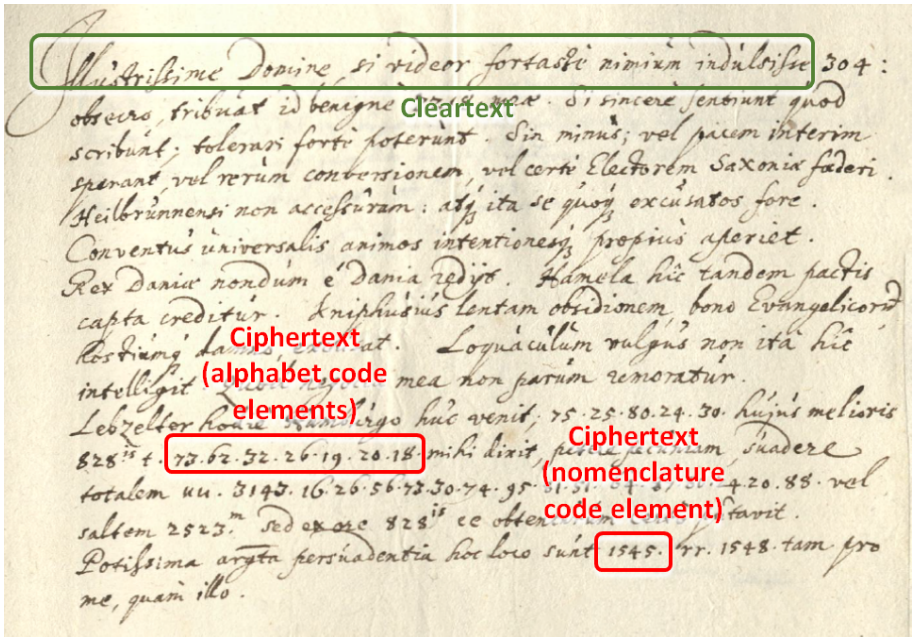

Figure 4: Example of a ciphertext: letter written by Adler Salvius to Axel Oxenstierna in 1633 using the key in Figure 3, Riksarkivet Sweden, Oxenstierna samlingen E 708:28.

Unambiguity Some terms, e.g. the word "code" have numerous meanings in various scientific fields, making it a source of confusion for readers. It is therefore recommended to avoid using it as a standalone term and instead provide more specific context or terminology to avoid misunderstandings. Thus, we try to avoid using terms that have various meanings in different disciplines.

Simplicity Our last, but nonetheless important goal was to make sure the text written using the proposed terminology is easily readable by people with various backgrounds.

3.2 Terminology usage

The full list of the proposed terms with their definitions is given in the Section 4.

To enable easier understanding of our proposal, most of these terms are illustrated in Figure 1. We will now explain how these terms are applied and how they relate to each other. A *cipher* is the algorithm used for encryption or decryption of information. The text which is meant to be encrypted is called a *plaintext*. The resulting encrypted text is known as *ciphertext*, which is made up of symbols from a *ciphertext alphabet*. Sometimes, an encrypted document may also contain non-encrypted text, known as a *cleartext*.

The process of encryption is controlled by the cipher *key*, and when the key is known, the ciphertext can be easily decrypted. Without the key, the process of analyzing the ciphertext to reveal the original plaintext is known as *cryptanalysis*.

Historical keys are typically composed of *plaintext elements* and their corresponding *code elements*. The plaintext elements are divided into two categories: alphabet elements (single letters) and the nomenclature elements (representing entities above the alphabet level). Similarly the code elements are composed of *alphabet code elements* and *nomenclature code elements*. The *nomenclature* is a part of a key which contains the nomenclature elements and their corresponding code elements.

Some keys also contain *empty code elements*, which are placeholders that can be filled in later, and *operational code elements*, which have special functions to carry out an operation on the revealed plaintext. Examples of operational code elements include *nulls*, which are fake code elements that encode an empty string in the plaintext, and *cancellation signs*, which mark the removal of a certain sequence of ciphertext. The relationship between different plaintext elements and code elements is shown in the Figure 2.

4 Proposed terms and their definitions

We propose to use the following terms:

Plaintext
: The text intended for encryption and/or the decrypted text.

Cleartext
: Intentionally unencrypted text in an encrypted document.

Ciphertext
: The encrypted text.

Encryption
: The process of transforming plaintext into ciphertext using a key.

Decryption
: The process of transforming ciphertext into plaintext using a key.

Cipher
: A set of rules (algorithm) describing the process of encryption/decryption.

Key
: A piece of information needed for encryption and decryption. A key has to be kept secret for security.

Cryptanalysis/Codebreaking
: The process of analyzing a ciphertext without knowing or partially knowing a key to reveal the original plaintext (and key).

Plaintext alphabet
: The set of elements used in the plaintext, e.g. letters, digits, punctuation marks, and spaces.

Ciphertext alphabet
: The set of symbols used in the ciphertext (e.g. digits, Latin and Greek letters, alchemical or Zodiac signs). We find these symbols not only in the ciphertext but also in the manuscript containing the key.

Plaintext element
: Any type of plaintext entity that has a corresponding code element assigned to it. It can represent a letter, double letter, syllable, name, function (e.g. preposition), or content word (e.g. noun, verb) as well as a phrase. The set of plaintext elements includes the alphabet and nomenclature elements.

Alphabet element
: Any letter in the alphabet of the writing system

that has a corresponding code element assigned to it. Alphabet elements constitute a subset of plaintext elements.

Nomenclature element
A plaintext element which is above the alphabet level. A nomenclature element can be a syllable, a name, a function and a content word as well as a phrase.

Code element
A symbol or a concatenation of symbols of the ciphertext alphabet used during the encryption for substitution of the corresponding plaintext element or to indicate that an operation on the revealed plaintext is needed. We distinguish between the following types of code elements: alphabet code elements, nomenclature code elements, and operational code elements.

Alphabet-code element
Code element used for encryption of one or several alphabet elements.

Nomenclature-code element
Code element used for encryption of a nomenclature element. Nomenclature elements are often encrypted using a different symbol type or of a different length than used for the alphabet code elements.

Nomenclature
A part of the key with a list of nomenclature elements and the corresponding nomenclature code elements.

Empty code element
Code element presented in the nomenclature which doesn't have any plaintext element assigned to it and is treated as a placeholder to be filled in later.

Operational code element
A code element that has a special function to carry out an operation on the revealed plaintext. Examples are repetition signs, cancellation signs, and nulls.

Repetition sign
An operational code element which indicates that the preceding letter in the revealed plaintext has to be repeated.

Cancellation sign/Nullifier
An operational code element which indicates that a certain sequence of a ciphertext (and hence the corresponding revealed plaintext) is to be removed.

Null/Nullity/Nullity sign/Blender
An operational code element which represents an empty string in the plaintext. Their purpose is to confuse the codebreaker or to mark the start and/or the end of the nomenclature elements.

Code separator / Token separator
A symbol or a concatenation of symbols that separates code elements or groups of code elements from each other. The main intention is to help the receiver to tokenize the ciphertext. In the case of cryptanalysis, it can help to break the cipher more easily.

5 Discussion

In this work, we discuss the terms which refer to cryptographic concepts that were actual before the 20th century when the widespread application of cipher machines began. Moreover, we are focused on the elements that are found in the ciphertexts and keys, which is only a part of the entire historical cryptography terminology.

We start by explaining some of the issues that we faced while designing our solution and provide the reasons for our decisions.

While working on the proposed terminology, we had to deal with trade-offs between perfect structure and simplicity. For instance, when referring to the elements on the ciphertext side it would be logical to use the term "ciphertext elements". Nevertheless, we use the term "code elements" which is commonly used in the area of historical cryptology. Moreover, it is shorter and easier to remember.

Another example is that we recommend using "nulls/nullities" without the word "sign" as it is also an already well-established term.

We also point out that "empty code elements" are not included in the set of "operational code elements". In fact, they do not indicate any operation, but rather refer to nomenclature code elements without a concrete plaintext element assigned yet.

Finally, we would like to mention that there is no strict border between the "alphabet elements" and "nomenclature elements." Often some nomenclature elements were presented in the same table as the alphabet elements. This is the reason why there is a curvy line between these two sets in Figure 2.

We now describe certain common cases where the inconsistent usage of terms may lead to confusion.

One of the most frequent examples is that in everyday life the word ”cipher” is used in the meaning of the ”ciphertext”. However, in scientific works, mixing these terms can lead to inconsistency or even misunderstanding. Hence, we would like to see this tradition be stopped.

Other terms that sometimes are used differently while other times as synonymous expressions are the ”nomenclature” vs ”nomenclator”. They might indicate a shorter or longer list of words with code elements, or the entire cipher key containing such a list.

The relation between the terms ”key” and ”cipher” may also become a source of confusion. In the scope of historical cryptology, these two words may sometimes be used with relatively close meanings. Given that for substitution ciphers, the key completely defines the concrete cipher, the evolution of key types also resulted in the parallel development of the corresponding ciphers. Nevertheless, the terms ”key” and ”cipher” have different meanings and should not be mixed.

Finally, we find the terms ”encipherment” and ”decipherment” problematic due to their ambiguous interpretations. For the term decipherment, the range of possible meanings varies from the synonym of ”decryption” which assumes the straightforward application of the cipher and the knowledge of the key, to ”cryptanalysis” where the plaintext and the key are revealed from the ciphertext. It may also mean an umbrella term for converting the ciphertext to the plaintext either using the key or by applying cryptanalysis. Additionally, the pair ”encipherment/decipherment” does not follow the symmetric principle, unless the ”decryption” is meant by ”decipherment”. Thus we believe that their usage may lead to confusion and the more specific terms decryption vs cryptanalysis would be preferable instead.

6 Summary

In this paper we presented a terminology for historical cryptology, focusing on the plaintext and the ciphertext sides of the encrypted sources. Our intention has been to contribute to a more systematic and consistent use of various terms, which we believe is especially important given the cross-disciplinary nature of our field. The terminology presented here is based on five principles: symmetry, expliciteness, hierarchy, unambiguity, and simplicity.

Our work presented in this paper shall be treated as just is: a work-in-progress and a point of departure to continue the discussion that started in 2018 in the HistoCrypt community. Noteworthy is that we only considered elements in the ciphertext and plaintext side, based on early modern ciphers. We have not suggested terminology for various types of operations, nor of cipher types, which would be the next step forward.

Acknowledgments

This work has been supported by the Swedish Research Council, grant 2018-06074, DECRYPT – Decryption of Historical Manuscripts.

References

Eugen Antal and Pavol Zajac. 2020. Hcportal overview. In *In Proceedings of the 3rd International Conference on Historical Cryptology. (HistoCrypt 2020)*, pages 18–20.

Eugen Antal. 2018. HC Portal. Portal of Historical Ciphers. `https://hcportal.eu/`.

Friedrich L Bauer. 1997. *Decrypted Secrets – Methods and Maxims of Cryptology*. Springer, 2 edition.

Elonka Dunin and Klaus Schmeh. 2020. *Codebreaking: A practical guide*. Hachette UK, London.

Employees of Bletchley Park. 1945. *A Cryptographic Dictionary*. NR 4559, Historic Cryptographic Collection, Pre-World War I Through World War II, Record Group 457, The National Archives and Records Administration (NARA) 8601 Adelphi Road, College Park, Maryland.

Bernhard Esslinger. 2023. *The CrypTool Book*. ArtechHouse.

William Frederick Friedman and Lambros D Callimahos. 1985. *Military Cryptanalytics*. Aegean Park Press.

William F Friedman. 1959. Military Cryptanalysis. *Aegean Park Press*.

David Kahn. 1996. *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*. Scribner, New York, NY.

Alan G Konheim. 1981. *Cryptography, a primer*. John Wiley & Sons, Inc.

George Lasry, Beáta Megyesi, and Nils Kopal. 2020. Deciphering Papal Ciphers from the 16th to the 18th Century. *Cryptologia*, pages 479–540.

Beáta Megyesi and Crina Tudor. 2021. Transcription of Historical Ciphers and Keys: Guidelines, version 2.0. https://cl.lingfil.uu.se/~bea/publ/transcription-guidelines-v2.pdf. Version: March 30, 2021.

Beáta Megyesi, Nils Blomqvist, and Eva Pettersson. 2019. The DECODE Database: Collection of Ciphers and Keys. In *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt19*, Mons, Belgium.

Beáta Megyesi, Crina Tudor, Benedek Láng, Anna Lehofer, Nils Kopal, Karl deLeeuw, and Michelle Waldispühl. 2022. Keys with nomenclatures in the early modern europe. *Cryptologia*.

Beáta Megyesi. 2020. Transcription of Historical Ciphers and Keys. In *Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt20*, Budapest, Hungary.

Aloys Meister. 1906. *Die Geheimschrift im Dienste der Päpstlichen Kurie von Ihren Anfängen bis zum Ende des XVI. Jahrhunderts*, volume 11. F. Schöningh.

Klaus Schmeh. 2018. Revisited: A terminology for codes and nomenclators. https://scienceblogs.de/klausis-krypto-kolumne/2018/10/07/revisited-a-terminology-for-codes-and-nomenclators.