What Was Encoded in Historical Cipher Keys in the Early Modern Era?

Beáta Megyesi and Crina Tudor
Uppsala University, Sweden
first.last@lingfil.uu.se

Benedek Láng and Anna Lehofer
BME/ELTE, Hungary
(lang|lehofer.anna)@filozofia.bme.hu

Nils Kopal
University of Siegen, Germany
nils.kopal@uni-siegen.de

Michelle Waldispühl
Gothenburg University, Sweden
michelle.waldispuhl@sprak.gu.se

Abstract

In this paper, we present an empirical study on plaintext entities in historical cipher keys from the 15th to the 18th century to shed light on what linguistic entities have been chosen for encryption. We focus mainly on the nomenclature part of the keys describing longer elements than the plaintext alphabet. We show that the chosen plaintext entities to be encoded varied over time. Nomenclatures developed from short lists consisting of names for persons and/or locations to longer, more advanced dictionaries and eventually to codebooks containing a highly diverse and advanced set of linguistic entities.

1 Introduction

The more experienced and knowledgeable person in historical cryptology would assume that cipher keys developed over time from simple to more sophisticated methods, such as from simple substitution to homophonic and polyphonic systems to polyalphabetic ciphers. Also, many of us would assume that the length of the keys developed over centuries from small keys with a few lines to codebooks with hundreds of pages. Such assumptions are often based on individual experiences. Large-scale studies on cipher keys have not been performed until recently.

In this paper, we study the development of cipher keys over time from one particular aspect, namely the plaintext elements in cipher keys. We investigate what was chosen to be encoded — apart from the alphabet — as part of the nomenclature. Inspired by linguistic knowledge about historical languages, we develop a taxonomy, a set of features with applied values to study plaintext entities within and across word boundaries. We then investigate how common each feature is in the keys and measure the occurrence throughout the centuries. In the following, we give an overview of previous work. In Section 3 we present the data and in Section 4 we describe how we annotated the keys. In Section 5 we describe our findings, and we discuss those in Section 6. Lastly, in Section 7 we conclude the paper.

2 Background

Cipher keys and nomenclatures have long been in the center of scholarly interest. Already in the late 19th and early 20th century representative collections of keys have been transcribed and published, albeit without any in depth analysis as far as their content is concerned (Rockinger, 1892; Meister, 1902; Meister, 1906; Devos, 1950). The next step was made by David Kahn in 1967, who in his grand oeuvre, The Codebreakers (Kahn, 1996), offered an extensive history of cryptography, and shared a lot of details about cipher keys. It was also Kahn, who — a few decades later — urged crypto-historians to carry out a systematic analysis of nomenclatures, as he put it: "timing and quantifying the change" (Kahn, 2008, p.58).

Several publications have been written on the content of cipher keys belonging to geographically well-defined areas: Benedek Láng on the early modern Central European area (Láng, 2018), George Lasry, Beáta Megyesi and Nils Kopal on the 16th to 18th century papal diplomacy (Lasry et al., 2020).

It was first by Megyesi, Tudor, Láng and Lehofer (2021) that a quantitative analysis was made on a European sample of 700 cipher keys although this sample was not representative for the whole of European history (Megyesi et al., 2021). This quantitative analysis was possible thanks to the DECODE database (Megyesi et al., 2019) which is aimed for the storage and description of historical encrypted sources and has be-

come the largest source for historical ciphers and keys by today. At the time of writing, the database contains ca 2900 sources: 1227 ciphers and 1665 original keys from various archives.

3 Data

For studying plaintext entities in keys, we use all keys that were available to us when we started the study in April 2021 through the DECODE database (Megyesi et al., 2019), i.e. 1610 key records in total. Out of all keys, 1384 contained nomenclatures with lists of plaintext entities beyond alphabet letters and were therefore chosen for further analysis. All keys are available through the database[1]. Throughout this paper we will refer to the record ID when we refer to specific records which are easily accessible by a simple search in the database.

The keys are kept in archives in 10 European countries: Austria, Belgium, France, Germany, Hungary, Italy, Spain, the Netherlands, UK, and the Vatican City State and are dated from between the 15th and the 18th centuries. The distribution across holder countries is presented in Figure 1. The data has been collected from 26 archives located in 10 different European countries. A large portion of the data comes from three Austrian archives, as well as five Hungarian archives and five Italian ones. For a complete list of the source holders, please consult Appendix A.

4 Annotation

All keys were manually annotated by the authors according to a linguistically motivated taxonomy developed for our purpose. Nine linguistic categories have been considered for annotation of each key including named entities, numbers, content and function words, syllables, morphological endings, phrases, sentences and punctuation marks. In the following, we list the single categories, explain them and give examples:

- *Named entities*: names and other forms of reference for persons, including titles and personal names (e.g. *Karl, the Emperor*), for places/locations (e.g. *Anglia*, *Belgium*) in Figure 2, or other kinds of entities (i.e. institutions, political entities, ships, etc.) as illustrated in Figure 3 .
- *Numbers*: any expression denoting numbers, written as words e.g. *centum* or *mille* in Figure 4, or as digits *1., 2., 3.* etc. in Figure 5.
- *Content words*: words that have meaning, including nouns (*general*), verbs (*gouverne*), adjectives or adverbs (*grand*), as illustrated in Figure 6.
- *Function words*: grammatical words that do not bear meaning by itself including prepositions, pronouns, conjunctions, auxiliary verbs, e.g. *par, pas, per, pour*, as shown in Figure 7.
- *Syllables*: a series of single, unbroken sound with a vowel and accompanying consonants that are not function words, e.g. *se, si, so, su* in Figure 8.
- *Morphological endings*: marking morpheme types, e.g. case endings as shown in Figure 9 with genitive (*genit.*), dative (*dativus*), accusative (*accus.*), and *ablative* (ablat.) listed with their respective codes. The same figure also states that the plural form of a word is created with the addition of an ”x”.
- *Phrases*: clusters of words that are neither titles nor in any other type of named entity, illustrated in Figure 10.
- *Sentences*: an independent clause that consists of a subject and a predicate, with or without other dependent clauses, e.g. ”All is well.” or the sentence in Figure 11.
- *Punctuation*: encoded punctuation marks, e.g. ”.”, ”?” or expressed in words such as ”coma” or ”punctum” as exemplified in Figure 12.

For annotation of the above mentioned features, first 30 then 350 keys were labeled by at least two annotators to discuss the annotation principles and set up guidelines, and later for consistency check to reach consensus among the group of annotators. Then, all remaining keys were annotated by one person. The annotators were encouraged to mark uncertainties with a question mark and put a note in the comment field about the particular difficulty. The entire group of annotators discussed those cases together and they were corrected later by another annotator.

[1] https://de-crypt.org/decrypt-web/RecordsList

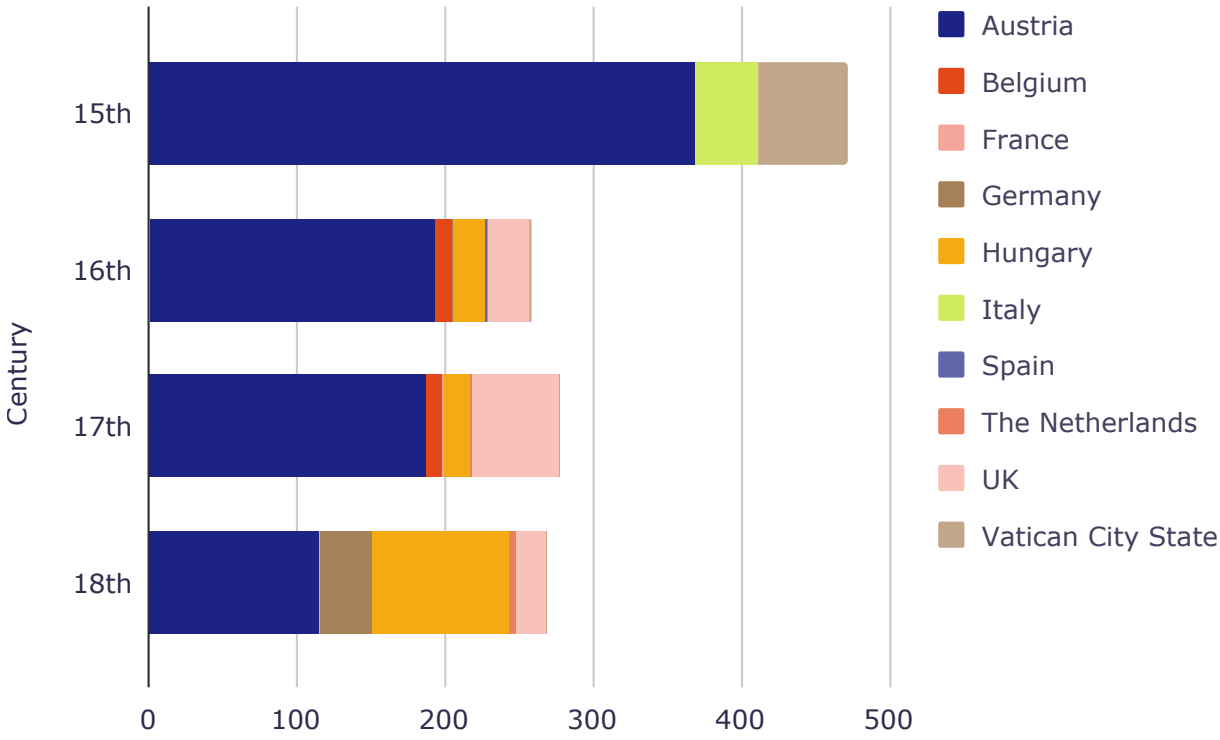

Figure 1: The distribution of keys across the holder countries.

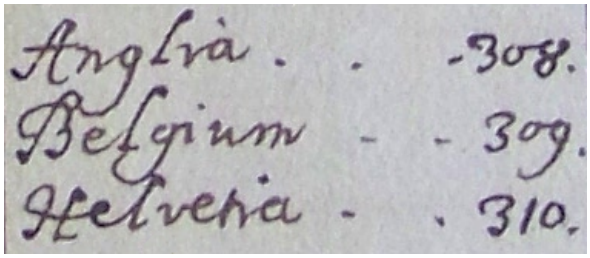

Figure 2: Example of named entities for places in nomenclatures from the early 18th century (Key 552, 1703).

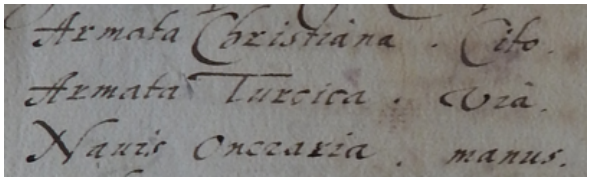

Figure 3: Example of other types of named entities such as armed forces or ships (Key 1317, 1579).

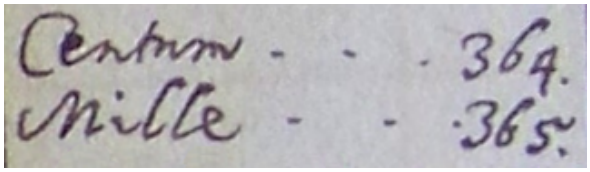

Figure 4: Example of numbers in nomenclatures from the early 18th century (Key 552, 1703).

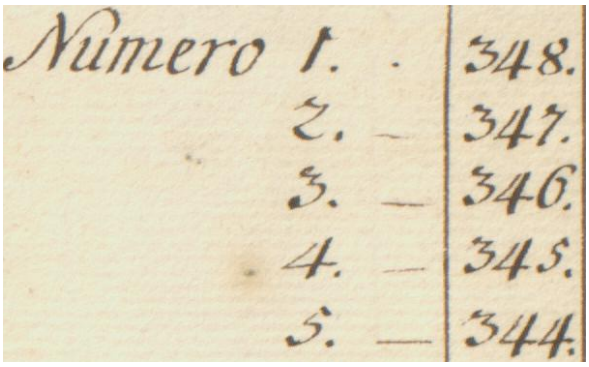

Figure 5: Example of numbers in nomenclatures (Key 2324, NA).

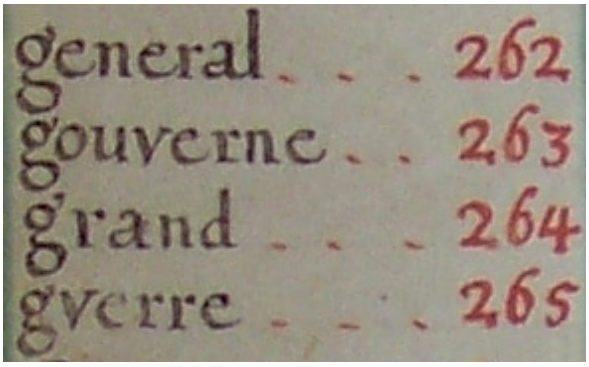

Figure 6: Example of content words in nomenclatures (Key 633, 1703).

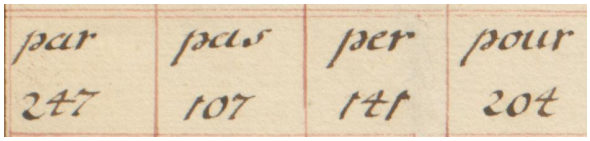

Figure 7: Example of function words in nomenclatures (Key 2330, NA).

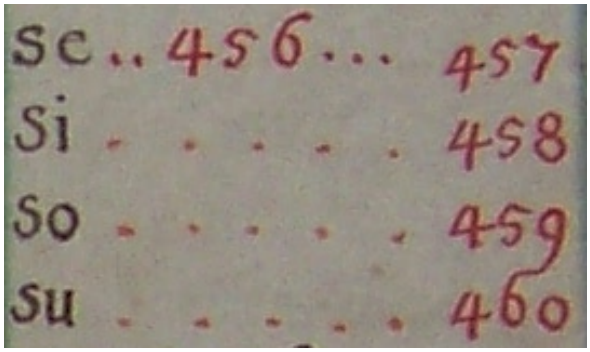

Figure 8: Example of syllables in nomenclatures from the early 18th century (Key 633, 1703).

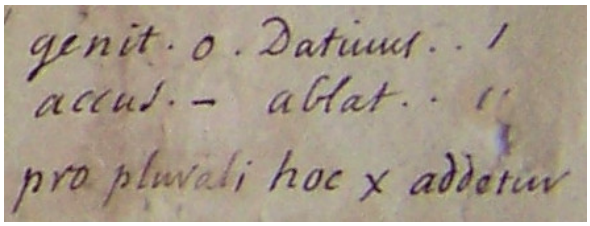

Figure 9: Example of morphological endings listed in a nomenclature (Key 673, 1660).

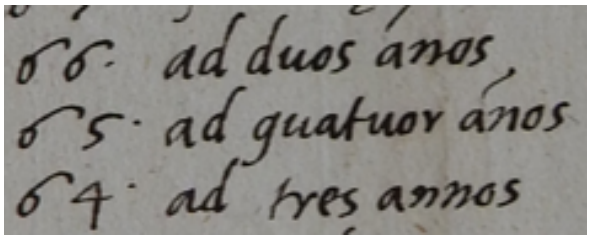

Figure 10: Example of phrases listed in a nomenclature from the 16th century (Key 1198, 1540).

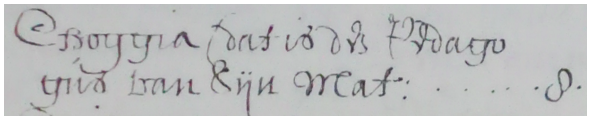

Figure 11: Example of a sentence listed in a nomenclature from 1620 (Key 2118, 1620).

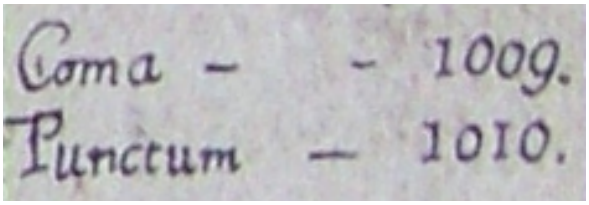

Figure 12: Example of punctuation in a nomenclature from the early 18th century (Key 616, 1703).

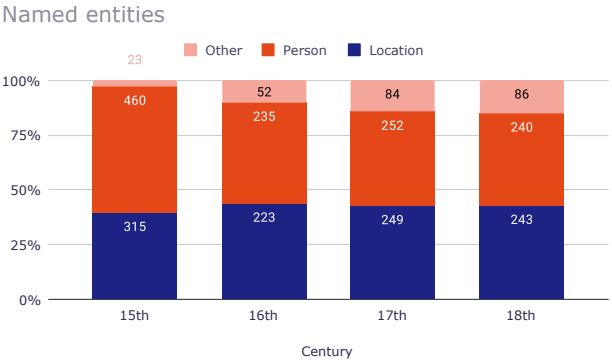

Figure 13: The percentage of keys with nomenclatures containing various types of named entities that include names for persons (P) in red, locations (L) in blue, and others (O) in pink.

5 Results

We summarize our findings per century for each plaintext type. We start with the occurrence of personal or place names, and other types of named entities. Names are generally a widespread plaintext type in nomenclatures. 96% of all keys analyzed in this study contain named entities. Figure 13 illustrates first the percentage of keys that include nomenclatures with persons, locations, and other types of referents throughout the 15th, 16th, 17th, and 18th centuries. Figure 14 presents the percentage of all possible combinations of named entity types. While the large number of nomenclatures contained persons and locations, other types of named entities became more frequent over time at the expense of entities referring to location. Other types could be armed forces or ships, as was exemplified in Figure 3.

The combination of named entities for persons (P) and locations (L) was the most frequent in all time periods. We also observe an increase of the combination of all three types (P, L, O) over time. Decreasing on the other hand is the occurrence of named entities for person (P) only while the type ”other” (O) as a single type in the nomenclature table increases slightly over time.

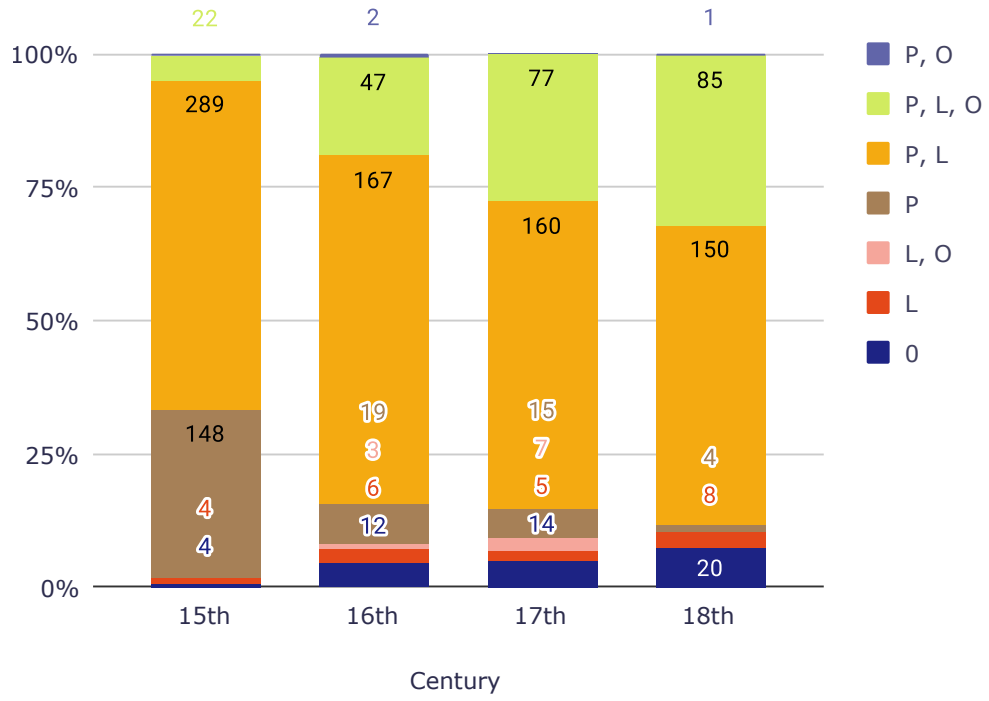

Figure 14: The percentage of keys with nomenclatures containing various types of named entities: named for persons (P), locations (L), and others (O).

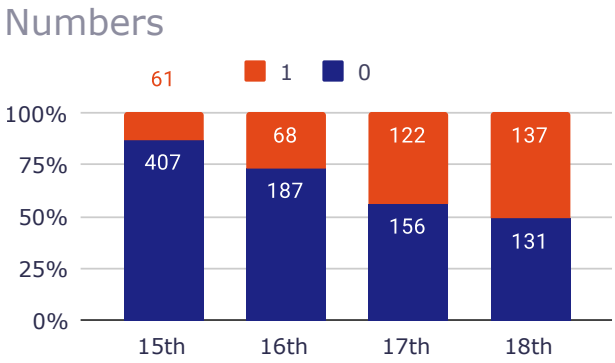

Figure 15: The percentage of keys with nomenclatures containing numbers marked as 1 in red.

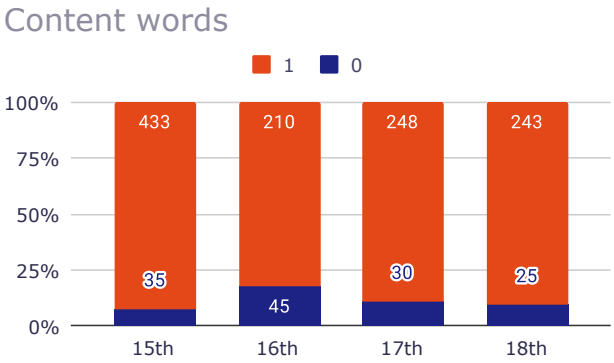

Figure 16: The percentage of keys with nomenclatures containing content words marked as 1 in red.

Nomenclatures might also include specific numbers, such as ”hundred”, ”thousand”, or ”million”. Listing numbers as part of the nomenclature increases greatly over time, as illustrated in Figure 15. In the 15th century, ca 20% of the keys contained numbers, while in the 18th century we find numbers as plaintext elements in more than 50% of the keys.

Content words have been always present throughout the centuries in some way or another, just as we could expect, as illustrated in Figure 16. However, the keys from the 16th century lack content words as part of their nomenclatures to a greater extent than during other time periods.

Similar to content words, function words were also often part of the nomenclature, but less frequently as content words, see the absolute numbers of occurrences in Figure 17. Interestingly, their presence decreased the most in the 16th century where less than half of the nomenclatures contained function words, and their presence increased again over time.

Listing syllables as part of the nomenclature became more common in the 17th century, and syllables are present in almost all keys with nomenclatures from the 18th century, as shown in Figure 18.

Morphological endings could also be encoded

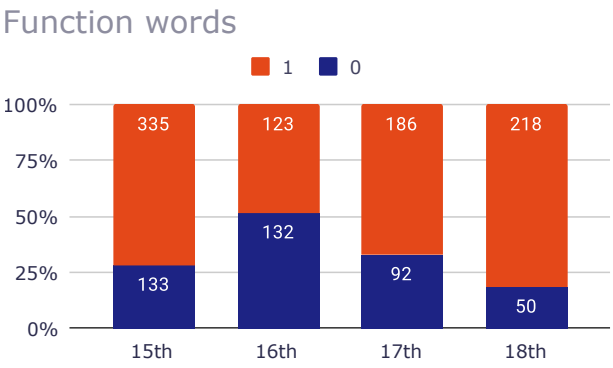

Figure 17: The percentage of keys with nomenclatures containing function words marked as 1 in red.

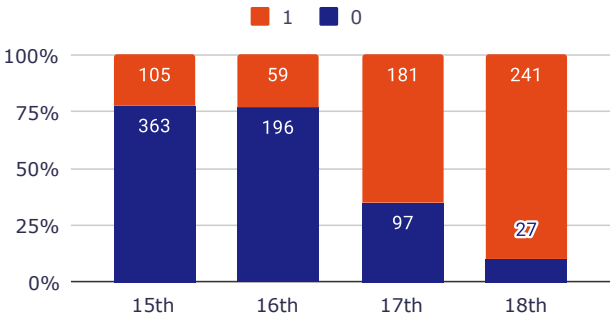

Figure 18: The percentage of keys with nomenclatures containing syllables marked as 1 in red.

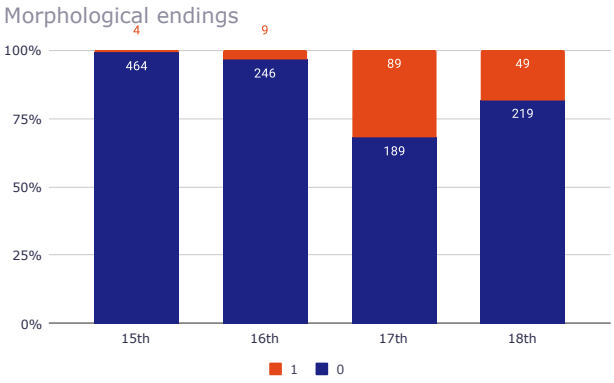

Figure 19: The percentage of keys with nomenclatures containing morphological endings marked as 1 in red.

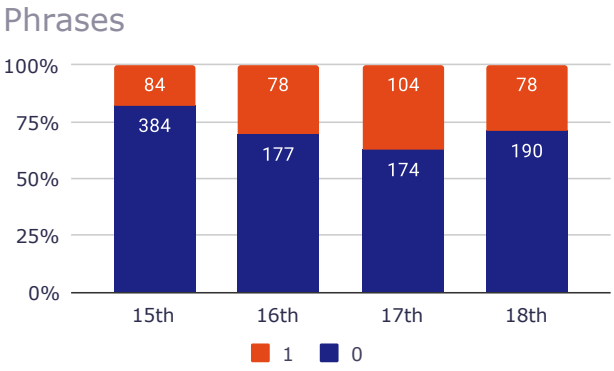

Figure 20: The percentage of keys with nomenclatures containing phrases marked as 1 in red.

and listed as part of the nomenclature. Encoding morphological endings as specific entities in nomenclatures was most common in the 17th century, to decrease in frequency later in the 18th century, as illustrated in Figure 19.

Using phrases became more common over time, but the majority of the keys lacked phrases, see Figure 20, with the exception of multi-word units of named entities, referring to persons or locations. We categorized the latter cases as named entities shown in Figures 13 and 14.

Encoding entire sentences in the nomenclature was hardly ever practiced during the centuries included in our sample, and the few that occurred were found in keys from the 16th and the 18th centuries, as shown in Figure 21.

Lastly, punctuation was normally not part of the nomenclature (neither of the alphabet) in the 15th century but became more frequent over time, as shown in Figure 22. This tendency might be due to the increased usage of these signs when writing conventions became more common in the 17th and

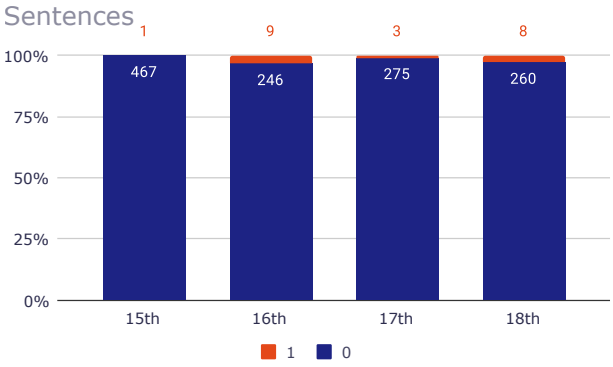

Figure 21: The percentage of keys with sentences in nomenclatures marked as 1 in red.

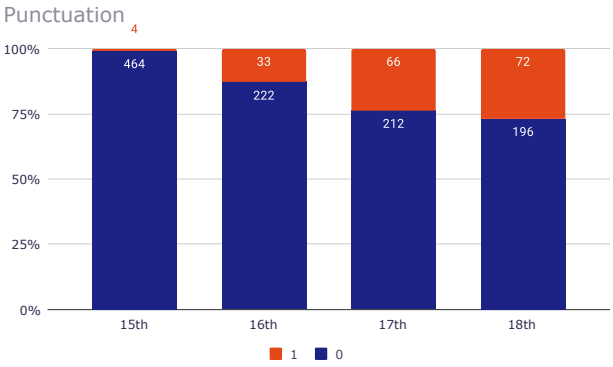

Figure 22: The percentage of keys with nomenclatures containing punctuation marks marked as 1 in red.

18th centuries.

6 Discussion

The results of investigating the plaintext elements in nomenclatures of 1384 keys show clear tendencies and evolution over time. The size of the nomenclatures in keys grows and the nomenclatures contain increasingly more plaintext entities on lower linguistic levels below words, such as morphemes and syllables. We find an increased number of nomenclatures containing content and function words, syllables, words denoting numbers, and punctuation marks in particular in the 17th and 18th centuries while named entities referring to persons and locations dominated in the 15th and 16th centuries. We can also see some tendencies that the keys from the 16th century became less advanced when compared to the 15th century material. In particular, the number of nomenclatures without content and function words (with the exception of names) increased in size in the 16th century to become frequent again later.

Although the data sample used in this study is unique in many ways — at least what the comparison with previous studies concerns — we are aware of that the set is not perfectly balanced, especially concerning the less optimal distribution across holder countries. We would like to include even more data from various states and regions used in internal and external communication in and between countries, institutions, and persons.

Also, we would have preferred to analyse the documents with regard to metadata over geographic areas where the cipher keys have been created and used rather than holder countries. Information about how and when particular keys were used, metadata about sender/receiver is missing for many records and could not therefore be analyzed.

In the future, we would like to extend our study not only to several keys from more diverse set of areas, but also to the code type and the encoding structure of various sorts of plaintext entities and correlate various features over time and over geographic areas. Analyzing the semantics of the named entities and the content words in detail would also be of high value to learn more about the specific vocabulary in nomenclatures from various time periods and areas, such as military terms, titles, or specific topics of interest. Manual analysis in such a detail with high precision would be highly time-consuming and expensive, and also error-prone with inconsistencies. Such an investigation could be done in large-scale by automatic methods only, albeit rather easily, if transcriptions of keys were available to a larger extent. We can only hope for the rapid development of image processing techniques in AI to serve our community with nice tools to be used for (semi-)automatic transcription.

7 Conclusion

In this study, we investigated the plaintext elements in over 1300 cipher keys containing nomenclatures. The keys were collected from ten European countries and originate from four centuries between 1400-1800. All of them were extracted from the DECODE database along with their metadata. We annotated manually all keys with respect to the plaintext elements in the nomenclatures looking at various linguistic features: the presence of named entities, numbers, content and function words, syllables, morphological endings, phrases, sentences, and punctuation marks.

We found that the content assigned with specific codes in the nomenclatures varied and evolved over time. The nomenclatures became longer with more varied vocabulary containing not only various types of named entities such as personal names and content words such as nouns, verbs or adjectives, but also function words, syllables and even morphological endings. We could also show that nomenclatures from the 16th century become less advanced in terms of what have been chosen to be encoded. This is supported by the fact that the number of keys with content and function words decreases in the 16th century compared to the 15th century to increase again from the 17th century and onwards. Encoding morphological endings can we expect in the keys from the 17th century, and syllables in keys from the 17th and 18th centuries. Including sequences of several words, i.e. phrases occurred throughout all times, most frequently in the 16th century, and punctuation marks can we expect in nomenclatures to some extent in the 16th century and more commonly from the 17th century. Our results are not surprising but we hope that they can be useful to make some assumption concerning what can be expected to be found in ciphertexts from various time periods in Europe.

Acknowledgments

This work has been supported by the Swedish Research Council, grant 2018-06074, DECRYPT – Decryption of Historical Manuscripts. We are very grateful to our colleagues in the project for useful comments and fruitful discussions, especially Karl de Leeuw and George Lasry. We would also like to thank Satoshi Tomokiyo and Anne-Simone Rous for their contributions to the DECODE database.

References

J. P. Devos. 1950. *Les chiffres de Philippe II (1555-1598) et du Despacho Universal durant le XVIIe siècle.* Brussels: Académie Royale de Belgique.

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

David Kahn. 2008. The future of the past—questions in cryptologic history. *Cryptologia*, 32:56–61.

Key 1198. 1540. Reproduced image from Österreichisches Staatsarchiv, Haus-, Hof- und Staatsarchiv, Staatskanzlei Interiora, Chiffrenschlüssel, Kt. 13. Fasc. 20. f 20., DECODE ID 1198, link: https://de-crypt.org/decrypt-web/RecordsView/1198.

Key 1317. 1579. Reproduced image from Österreichisches Staatsarchiv, Haus-, Hof- und Staatsarchiv, Staatskanzlei Interiora, Chiffrenschlüssel, Kt. 13. Fasc. 20. f 244-245., DECODE ID 1317, link: https://de-crypt.org/decrypt-web/RecordsView/1317.

Key 2118. 1620. Reproduced image from Nationaal Archief, 1.01.02 (Staten-Generaal), inventaris nummer 689, Cornelis Haga, DECODE ID 2118, link: https://de-crypt.org/decrypt-web/RecordsView/2118.

Key 2324. N/A. Reproduced image from Hessisches Staatsarchiv Darmstadt (HStAD), 10024, Loc. 8236/11, f. 111, DECODE ID 2324, link: https://de-crypt.org/decrypt-web/RecordsView/2324.

Key 2330. N/A. Reproduced image from Hessisches Staatsarchiv Darmstadt (HStAD), 10024, Loc. 8236/11, f. 121, Great Cipher of Saxony, DECODE ID 2330, link: https://de-crypt.org/decrypt-web/RecordsView/2330.

Key 552. 1703. Reproduced image from the National Archives of Hungary, G15 Caps. C. Fasc. 43. 3., DECODE ID 552, link: https://de-crypt.org/decrypt-web/RecordsView/552.

Key 616. 1703. Reproduced image from the National Archives of Hungary, G15 Caps. C. Fasc. 43. 62., DECODE ID 616, link: https://de-crypt.org/decrypt-web/RecordsView/616.

Key 633. 1703. Reproduced image from the National Archives of Hungary, G15 Caps. C. Fasc. 44. 01., DECODE ID 633, link: https://de-crypt.org/decrypt-web/RecordsView/633.

Key 673. 1660. Reproduced image from the National Archives of Hungary, P1238 Mihály Teleki Collection. Miscellaneous documents. Cipher keys. 13., DECODE ID 673, link: https://de-crypt.org/decrypt-web/RecordsView/673.

Benedek Láng. 2018. *Real Life Cryptology: Ciphers and Secrets in early modern Hungary*. Atlantis Press, Amsterdam University Press.

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

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, June.

Beáta Megyesi, Crina Tudor, Benedek Láng, and Anna Lehofer. 2021. Key Design in the Early Modern Era in Europe. In *Proceedings of the 4th International Conference on Historical Cryptology, HistoCrypt21*.

Aloys Meister. 1902. *Die Anfänge der modernen diplomatischen Geheimschrift.* Paderborn: Ferdinand Schöningh.

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

Ludwig von Rockinger. 1892. Über eine bayerische Sammlung von Schlüsseln zu Geheimschriften des sechzehnten Jahrhunderts. *Archivalische Zeitschrift*, pages 18–92.

Appendix A. List of archives

- Austria, Vienna

  - Österreichische Nationalbibliothek

  - Österreichisches Staatsarchiv, Haus-, Hof- und Staatsarchiv

- Belgium, Brussels

  - Algemeen Rijksarchief (State Archives)

- France, Paris

  - Bibliothèque nationale de France (BNF)

- Germany

  - Politisches Archiv des Auswärtigen Amtes (PAAA) (Berlin)

  - Saxon State Archive (Dresden)

  - German Federal Archives (Koblenz)

- Hungary, Budapest,

  - Magyar Tudományos Akadémia, Kézirattár (Hungarian Academy of Sciences, Department of Manuscripts)

  - Hadtörténeti Levéltár (Military History Archive)

  - Magyar Nemzeti Levéltár Országos Levéltára (National Archives of Hungary)

  - Országos Széchényi Könyvtár (National Széchényi Library)

  - Ráday Levéltár (Ráday Archives)

- Italy

  - Archivio di Stato di Firenze (Florence)

  - Archivio di Stato Lucca (Lucca)

  - Matova State Archive (Mantova)

  - Archivio di Stato di Milano (Milan)

  - State Archives of Venice (Venice)

- Spain, Madrid

  - Biblioteca Nacional (BNE)

- The Netherlands, The Hague

  - Koninklijk Huisarchief (KHA) (Royal Household Archives)

  - The Hague, Museum voor Communicatie (Museum of Communication)

  - The Hague, Nationaal Archief (National Archives)

- UK, Kew

  - The National Archives

- Vatican City State, Vatican City

  - Biblioteca Apostolica Vaticana

  - Archivio Apostolico Vaticano