# Deciphering Historical Syllabic Ciphers

**George Lasry**
The DECRYPT and CrypTool Projects
george.lasry@gmail.com

## Abstract

Historical ciphers with syllabic elements are significantly more challenging for cryptanalysis than regular homophonic ciphers. We present here a novel computerized technique which recovers significant parts of the keys, allowing for the remaining parts to be manually completed. We solved several previously undeciphered French, Spanish, and Italian syllabic ciphers, and we also evaluated the performance of this method against a series of additional historical syllabic ciphers.

## 1 Introduction

The ciphers used in Europe from the 15th century and until the 18th century were primarily homophonic, with a nomenclature of varying size. In recent years, several computerized techniques have been developed and successfully applied to the deciphering of historical documents encrypted with homophonic ciphers. Those techniques, however, are ineffective against syllabic ciphers. In this article, we describe various types of syllabic ciphers in Section 2, and the challenges in deciphering them in Section 3. In Section 4, we present the new algorithm which can recover significant parts of the key. With this initial partial solution, a cryptanalyst familiar with the language can easily recover most of the remaining key elements. In Section 5, we provide several case studies of successful decipherment. In Section 6, we evaluate the performance of the algorithm against several historical syllabic ciphers. We conclude our results in Section 7.

## 2 Syllabic ciphers

Homophonic ciphers consist of a list of symbols representing letters of the alphabet – more than one per letter, as well as a nomenclature with symbols representing common words, persons, places, punctuation signs, signs for doubling consonants, for repeating, or for deleting the previous symbol, or nulls. Syllabic ciphers are an extension of homophonic ciphers, adding dedicated symbols to represent various types of syllables, such as: [1]

- **Consonant-vowel (CV)** syllables, such as MA/ME/MI/MO/MU or TA/TE/TI/TO/TU.
- **Vowel-consonant (VC)** syllables, such as EB/EC/ED/EF etc.
- **Consonant-consonant-vowel (CCV)** syllables, such as PRA/PRE/PRI/PRO/PRU.
- **Consonant-vowel-consonant (CVC)** syllables**,** such as PAR/PER/PIR/POR/PUR.

We refer to the letters of the alphabet, the syllables, and the words, persons, places which are part of the nomenclature as the **plaintext vocabulary**. With homophonic ciphers, there was usually only one way to decompose a given word before encryption, into plaintext vocabulary elements. We illustrate this with the English the word ESTABLISHED, which first needs to be decomposed into E-S-T-A-B-L-I-S-H-E-D, then enciphered.

With a cipher with CV syllables, there are additional options, such as E-S-**TA**-B-**LI**-S-**HE**-D, E-S-**TA**-B-**LI**-S-H-E-D, or E-S-**TA**-B-L-I-S-H-E-D.

With a cipher with also VC syllables, we have additional options, such as **ES**-**TA**-B-**LI**-S-H-**ED.**

With more complex syllables (e.g., CCV), we have even more options, such as E-**STA**-**BLI**-S-**HE**-D.

[1] In some cases, VCC syllables were encoded, e.g., EST, as well as elements composed only of consonants, such as TR or STR.

In the ciphers we examined, we saw two main patterns of how words are decomposed before encryption:

- **Random decomposition**: Any of the options for decomposing a word could be used, and often, the same word can be decomposed differently within the same ciphertext.
- **Systematic decomposition**: Decomposition is systematic and predictable, as described below.

We illustrate the case of systematic decomposition with the word ESTABLISHED, assuming the cipher has CV, VC, and CCV syllables.

- There are two options to start, E and ES. We select the longest one, ES.
- Next, we decompose TABLISHED. There are two options to continue, T or TA. Again, we select the longest, TA.
- Next, we decompose BLISHED. There are two options, B and BLI. We select the longest one, BLI.
- Similarly, we decompose the rest, obtaining **ES-TA-BLI-SHE-D**.

In some historical ciphers, we also see syllables spanning two adjacent words, for illustration purposes, the expression AN IDEA can be decomposed as A-NI-DE-A, the syllable NI including the last letter of the first word and the first letter of the second word.

The set of symbols for syllabic ciphers is usually significantly larger than for homophonic ciphers. For example, to represent all the CV syllables, with 17 consonants (B, C, D, F, G, H, L, M, N, O, P, QU, R, S, T, V, Z) and the 5 vowels, 85 additional symbols are needed. A similar number is required to represent VC syllables. For CCV and CVC syllables, dozens of additional symbols were needed.

For that purpose, instead of adding totally new symbols into the cipher key tables, diacritics were employed to alter the meaning of other symbols. For example, if the numerical symbol ***36*** represents the letter T, ***36:*** (***36*** with a colon on the right) could represent the syllable TA. Similarly, ***36.*** could represent the syllable TE, etc.[2] Such diacritics were added either on the top, bottom, left or right side of the symbol. Furthermore, two diacritics could be added, usually to represent CCV or CVC syllables. For example, ***36:'*** would represent TRA (the added ***'*** means that the letter R should be inserted between T and A). If diacritics are used consistently, e.g., ***46:*** is CA, ***46'*** is CE, we denote such sets of syllabic symbols as **regular syllabic symbols**.

There were cases, which we denote as **irregular syllabic symbols**, in which diacritics were not employed in a systematic manner. For example, ***36:*** means TA, but ***46'*** means CA (rather than ***46:***). Furthermore, in fully irregular syllabic ciphers, ***36:*** could be TA, but ***52'***, with a diacritic added to another unrelated numerical code (***52***), would represent TE. There were also cases in-between, with partial regularities.

Ciphers with regular syllabic symbols are significantly easier to solve. For example, if we know that ***36:*** represents TA, and that ***46.*** represents CI, then ***36.*** is likely to represent TI. In the algorithm we present in this article, we did not take advantage of such regularities in some ciphers, as we wanted to implement a solution applicable to the more general case.

This description is not comprehensive as we did not conduct a systematic survey of historical syllabic ciphers, and this paper is instead focusing on cryptanalysis techniques. The sample syllabic ciphers we analyzed are from Italy (15th and 16th centuries), Spain (16th and 17th centuries), and France (17th and 19th centuries).

## 3 Cryptanalytic challenges

The cryptanalysis of syllabic historical ciphers is significantly more challenging than of regular homophonic ciphers, such as a much larger key space. The set of cipher symbols most often consists of a few hundred distinct symbols. The size of key space is exponentially related to this number. Also, it may not be practical to compute n-gram statistics for very large vocabularies.

The types of syllables (CV, VC, CCV, CVC) the cipher employs and the decomposition scheme may vary and are generally unknown upfront.

[2] In some cases, diacritics were added to a new symbol, rather than to the one representing the base letter. For example, ***36*** could be T, but TA would be ***72:*** rather than ***36:***.

Existing computerized codebreaking algorithms for regular (non-syllabic) homophonic ciphers could only provide the meaning of most of the letter homophones, and a manual process was most often required to interpret the remaining symbols. Given the additional challenges with syllabic ciphers, our goal was to develop an algorithm that would be able to recover enough parts of the key and the plaintext, so that the remaining parts of the key may be reconstructed, and the ciphertext fully decrypted, with manual interactive work by a cryptanalyst.

## 4 The codebreaking algorithm

The algorithm is an extension, with substantial adaptations, of a simulated-annealing algorithm developed to solve homophonic ciphers (Kopal, 2019).

Due to the size of both the key space and of the plaintext vocabulary, the algorithm typically requires **extensive computing power**, e.g., a computer with dozens of cores, or multiple computers, running parallel instances of simulated annealing, to obtain useful results in minutes rather than in hours. In our tests, we ran the algorithm on a 64-core Windows 10 Pro PC with 256Gbytes of RAM memory. The algorithm may also require extensive trial-and-error to fine tune its parameters, which include:

- The expected pre-encryption word **decomposition scheme**. The algorithm we developed supports two schemes, **systematic** (deterministic) and **random**.
- **The set of syllables** expected: **CV, VC, CCV, CVC**, or any combination thereof.
- The **maximum number of homophones** per type of vocabulary element: Per vowel, per consonant, and per each type of syllables.
- Specifying letters that are **interchangeable**, e.g., U and V in French, Spanish, and Italian, or I, J, and Y in French.
- Letters that should be **replaced** with other letters (e.g., K with C, W with V), or **ignored** (e.g., X or Y in Italian).
- Whether **repeated letters** (e.g., LL, SS, TT) are represented by dedicated symbols.
- **A set of reference texts** of reference texts in the expected language, to compute n-gram statistics. While the algorithm is language-agnostic, some assumptions must be made on the plaintext vocabulary and the decomposition scheme before creating a database of n-gram statistics, which will be based on sequences of elements in the expected vocabulary, rather than just letter n-grams. These assumptions are formulated using the previously listed parameters. In addition to counting the occurrences of n-grams of single letters like E-S-T, S-T-A, or T-A-B, we also must count n-grams such as ES-TA-B or E-STA-B. As a result, n-gram statistics must be computed ad-hoc based on specific vocabulary parameters, rather than relying on pre-computed statistics.
- The **n-gram size**: 4-grams were empirically found to be the most effective, while 3-grams or 5-grams might be useful in some cases.

Other parameters are optional and can help the algorithm to converge better and faster:

- A small **set of common words** expected to be found in the nomenclature.
- Some **limitations** on the set of symbols allocated to letter homophones, such as allowing only symbols without diacritics, or numerical codes within a certain range, to be assigned as letter homophones.
- The **maximum number of distinct ciphertext symbols** to be considered. The algorithm discards the less frequent ones. This effectively reduces the size of the search key space.
- **Tentative key assignments** of symbols to vocabulary elements, as they are being identified with the semi-automated work described later in this section. If correct, those allow simulated annealing to produce a better and more complete solution, quickly and more reliably.

Simulated annealing starts by randomly allocating ciphertext symbols to plaintext vocabulary elements, prioritizing the most frequent ones. This means that the less frequent vocabulary elements will be ignored if there are not enough distinct ciphertext symbols, or if the number of processed ciphertext symbols has been limited.[3]

During simulated annealing, the only allowed key change is swapping the assignment of any two cipher symbols (for example, if ***32:*** was assigned to TA, and ***43'*** was assigned to CE, after the swap, ***32:*** is assigned to CE, and ***43'*** is assigned to TA). As a result, and in contrast with other solvers of homophonic ciphers, the number of symbols allocated to each vocabulary element is constant. This was found to provide for a more stable and more effective algorithm.

A score measuring the quality of the deciphered is computed as follows:

- Before starting simulated annealing:
    - Parse all the reference plaintexts, decomposing words into the specified vocabulary elements, according to the specified decomposition scheme.
    - Compute $F_g$, the relative frequencies for every combination $g$ of four successive elements of the vocabulary, a.k.a. *4-gram*, such as E-S-T-A, or CO-N-TRO-L, which appear in reference plaintexts.[4]

- During the search with simulated annealing, evaluate a candidate key as follows:
    - Decipher the ciphertext using the candidate key.
    - Compute $N_g$, the number of occurrences in the decrypted text of each 4-gram $g$.
    - Compute $N_c$, the number of occurrences in the decrypted text of each vocabulary element $c$.
    - The score $S$ for the tentative decipherment is computed as follows:

$$S = \sum_g N_g \log F_g / \sum_c N_c^2$$

When the algorithm starts producing tentative decryptions, it also highlights (in capital letters) plausible segments of vocabulary elements, if those can be found in the specified reference texts. This is especially useful if the cryptanalyst is not familiar with the plaintext language. It is expected that as more elements of the key are correctly recovered, there will be more of those highlighted plausible segments. Furthermore, the algorithm counts how many times each symbol occurs in such a plausible segment, and those occurring dozens of times are listed as likely to have been correctly assigned.

Working with the tool is done iteratively. At first, the program may produce only a few highlighted segments, and a few key assignments suggestions. The cryptanalyst manually reviews those segments and suggestions, and if they look plausible, they can be entered as parameters for the next run.

An example of an initial run is given in Figure 1. It is possible to discern several highlighted plausible French words or expressions, such as EN PRENDRE, GRANDEMENT LE, and IMPRIMER LE. The correct assignment of several ciphertext symbols (e.g., those representing N, E, T, R) composing those expressions are also be validated by the statistics listed below the decrypted text, and those symbols may be safely assigned accordingly for subsequent runs, by setting the tentative key assignments' parameter.

Running the algorithm again with the revised parameters will reveal additional assignments. Non-highlighted segments that are plausible may also reveal additional plausible assignments. When enough elements have been recovered, it may neither be necessary nor useful to run the automated algorithm again and the remaining work can be completed manually.

[3] By setting the parameters which specifies the maximum number of distinct ciphertext symbols to be processed.

[4] Or 3-grams, or 5-grams.

```
b-o< 14 U^ L S H :- S -8 r& S i- r& b-o_ 18 :- O. S d& 20 d. b U_ S u&
sv   h  ti c N T E  N P  RE N D  RE SO   vv e  me n ce z  ca r so n de
svhticNTENPRENDRESOvvemencezcarsonde

D< a b-o& a -i oio \o r. S u& n& S H U& + L 6 : ii r& + m_ i= b n. g
m  i se   i a  y   G  RA N DE ME N T LE a c f d i  re a po v  r ma l
miseiayGRANDEMENTLEacfdireapovrmal

ii ap -8 r^ n& b U& + n^ O^ + H r& + u& !  x^ b-o& S b-o_ S H :- ap m& +
I  M  P  RI ME R LE a mi ni a t re a de rr i  se   N SO   N T E  M  PE a
IMPRIMERLEaminiatreaderriseNSONTEMPEa

135 times: S    -> N
 89 times: t&   -> E
 79 times: H    -> T
 75 times: b    -> R
 60 times: i=   -> V
 36 times: :-   -> E
 35 times: +    -> A
 35 times: r&   -> RE
 33 times: a    -> I
 28 times: u&   -> DE
 23 times: =    -> O
 22 times: U&   -> LE
```

Figure 1– Sample printout of the algorithm for syllabic ciphers

## 5 Decipherments

With this technique and the semi-automatic process described in Section 4, we deciphered several documents encrypted with historical syllabic ciphers, for which the key and the plaintext were not known in advance.

### 5.1 Archivio di Stato di Milano - Visconteo Sforzesco Segnatura

A letter in Italian from Ottone de Carretto from 1457. We analyzed about 3,900 ciphertext symbols, with 113 distinct ones. The cipher features symbols for VC syllables assigned in an irregular manner, and words are decomposed randomly. The reconstructed key is shown in Appendix 1. It later turned out that a copy of the cipher key is held in the Archivio di Stato di Milano.[5]

### 5.2 Simancas EST LEG 1381 – 143

A letter in French, from 24 August 1551. After deciphering the letter, it turned out to be using the same cipher used between Charles V and his ambassador Jean de Saint-Mauris (Pierrot et al., 2023). We analyzed a total of 4,300 ciphertext symbols. There are 148 unique cipher symbols, some with diacritics to represent CV syllables and a few CCV syllables, assigned in a regular manner. Word decomposition is random.

### 5.3 Simancas EST LEG 1381 - 180

A letter in Spanish from 15 September 1551.[6] It contains about 2,300 ciphertext symbols, with 123 distinct ones. It features CV syllables, as well as a few CCV syllables, marked with diacritics, and assigned in a regular manner. Word decomposition is systematic.

### 5.4 BnF Clairembault 421 f. 160

A letter from Henri Brasset, a French resident in the Hague, to Cardinal Mazarin, at the time the chief minister of infant King Louis XIV, written on 30 March 1649 and held in the Bibliothèque Nationale de France.[7] The ciphertext has about 1,800 ciphertext symbols, and 156 unique symbol types. Some archive images are damaged, several parts missing or illegible. Non-numerical symbols represent letter homophones. The other elements (CV syllables, nomenclature) use two-digit numerical codes with optional diacritics, but the syllable symbols are mostly assigned in an irregular manner. Word decomposition was random. Overall, cryptanalysis was quite

[5] ASMi Carteggio Visconteo Sforzesco Segnatura1598 f.89.

[6] The decipherment includes some indications on the possible sender and recipient: "Copiado loque Su Magestad scrive a [Principe] Doria a v de setienbre presinte", "Al seno Ferando", "Al enbaxador Figueroa." More details on this cipher and other Spanish syllabic ciphers in (Tomokiyo, 2023, direct link).

[7] This cipher was first presented as an unsolved cipher by Satoshi Tomokiyo in (Tomokiyo 2023, direct link).

challenging. After decrypting the ciphertext, we were able to find the original plaintext in French archives,[8] and to complete most of the key assignments, as shown in Appendix 2.

### 5.5 Dresden - Militärhistorisches Museum

An unpublished letter, recently discovered in the archives of Russian Field Marshall Michail Andreas Barclay de Tolly. The letter was written on 4 August 1813 by General Rapp to Napoleon's headquarters, during the siege of Danzig. The cipher consists of about 900 ciphertext symbols, with 109 unique symbols, composed of one or more digits, mostly in the range up to 200, without any diacritics. After cryptanalysis, we established that the cipher features CV syllables, assigned irregularly, and that words are decomposed randomly. It later turned out that the cipher key was a known version of Napoleon's Small Cipher.[9]

## 6 Performance evaluation

We also analyzed the performance of the algorithm against additional syllabic ciphers for which the key was already known.[10]

- **Baldassare Castiglione to Niccolò Schomberg**, 25 March and 3 April 1527.
- **Simancas EST LEG 1386 - 1**. From 15 March 1577, from Pedro Gonzalez de Mendoza to King Philip II.
- **ARA Brussels SEG 2559.** A series of letters in Spanish from 1674-1678 also sent to Balthazar de Fuenmayor.
- **ARA Brussels SEG 2559.** A letter in French, 31 July 1676, sent to Balthazar de Fuenmayor, Spanish ambassador in Denmark.
- **KHA Amsterdam, Willem II/XIII-I**. A letter, from 9 January 1684 from le Comte d'Avaux, the French ambassador in Holland, to King Louis XIV.

In Figure 2, we summarize the performance of the new algorithm, tested against those ciphers and the other five listed in Section 5. The accuracy numbers refer to the **decrypted text accuracy** – the percentage of the ciphertext symbols in the documents correctly decrypted, and the **reconstructed key accuracy** – the percentage of the symbol types (distinct ciphertext symbols) correctly assigned. The accuracy of the decrypted text is always higher than the accuracy of the reconstructed key, as lower frequency symbol types are often ignored by the algorithm (and therefore, not assigned, thus reducing the key accuracy), and sparsely used symbols are more likely to be incorrectly interpreted. More importantly, the accuracy of the decrypted text is the main factor affecting the ability to make further progress manually.

Achieving those promising performance numbers with the automated algorithm required extensive trial-and-error and tweaking of the parameters, especially for those ciphers for which the types of syllables used were not known. The algorithm turned out to be highly sensitive to the parameters specifying the maximum number of homophones per vowel or per consonant, and the word decomposition scheme – a wrong selection would often prevent the algorithm from converging.

We also tested the performance of the algorithm when limiting the number of ciphertext symbols to 1,000, so that the performance may be compared across the various ciphers, as shown in the two rightmost columns. The performance is strongly affected by the number of distinct symbol types, especially for ciphers with more than 150 symbol types, which require longer ciphertexts to obtain satisfactory results. In general, an initial accuracy of 40% or above is most often enough to decipher a ciphertext with the semi-automated process described in this Section 4.

[8] BnF Français 17901 f.230.

[9] More details in (Tomokiyo, 2023, direct link).

[10] We recovered the keys for the first two items based on plaintext inscribed on the margins. The key for the Comte d"Avaux cipher was recovered based on a similar key (Lasry, 2019). The key for the fourth item was recovered by Carlos Köpte (Tomokiyo, 2023), and the key for the fifth one by Norbert Biermann (Simonetta, 2023).

| Reference and origin | Language and year | Length | Symbol types | Syllables | Regular syllables | Word decomposition | Accuracy - Decrypted text and key | | Accuracy with only 1000 symbols | |
|---|---|---|---|---|---|---|---|---|---|---|
| **ASM – Ottone de Carretto** | Italian 1457 | 3,900s | 113 | VC | No | Systematic | **75%** | 58% | **27%** | 13% |
| **Castiglione to Schomberg** | Italian 1527 | 900 | 134 | CV | Partially | Systematic | **35%** | 18% | | |
| **Simancas EST LEG 1381 180** | Spanish 1551 | 2,300 | 123 | CV CCV | Yes | Systematic | **78%** | 51% | **49%** | 30% |
| **Simancas EST LEG 1381 143** | French 1551 | 4,300 | 148 | CV CCV | Yes | Random | **81%** | 44% | **<5%** | <5% |
| **Simancas EST LEG 1386 1** | French 1577 | 1,200 | 156 | CV VC CCV | Yes | Systematic | **78%** | 47% | **<5%** | <5% |
| **BnF Clairembault 421 f. 160** | French 1649 | 1,800 | 156 | CV | Yes | Random | **80%** | 48% | **<5%** | <5% |
| **ARA SEG 2559 - French** | French 1674 | 2,300 | 101 | CV | No | Random | **96%** | 68% | **86%** | 51% |
| **ARA SEG 2559 - Spanish** | Spanish 1674-1678 | 3,200 | 178 | CV | Yes | Systematic | **73%** | 33% | **35%** | 16% |
| **KHA Prins Willem II/XIII-I** | French 1684 | 4,200 | 219 | CV | Partially | Random | **78%** | 39% | **<5%** | <5% |
| **Dresden** | French 1813 | 900 | 109 | CV | No | Random | **76%** | 50% | | |

Figure 2 – Performance evaluation of the new codebreaking algorithm for syllabic ciphers

## 7 Conclusion

Before this work, the cryptanalysis of a syllabic cipher was highly challenging and most often possible only if the structure of the syllabic symbols was regular, or if a matching plaintext could be found. The new algorithm presented here can provide an initial breakthrough, which, with manual analysis, can most often lead to a successful decipherment of historical syllabic ciphers, as exemplified here.

In addition, the algorithm was designed for the general and more challenging case of irregular syllable symbols. It may be easily improved to take advantage of possible regularities in the assignment of the ciphertext symbols to syllables.


## Acknowledgments

The author would like to thank Jessika Novak for providing a copy of the Ottone de Carretto cipher, Klaus Schmeh, Wolfgang Schmidt, and Erik Zimmermann for publishing the Dresden cipher, Satoshi Tomokiyo for bringing the Brasset-Mazarin cipher to light and for reviewing an early version of the paper, Paolo Bonavoglia for providing a copy of original the Ottone de Carretto cipher key and helping with the decipherment, and Norbert Biermann for reviewing the paper and helping to fine-tune decipherments.

## Funding

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


## References


Nils Kopal, 2019. *Cryptanalysis of homophonic substitution ciphers using simulated annealing with fixed temperature*. HistoCrypt 2019.

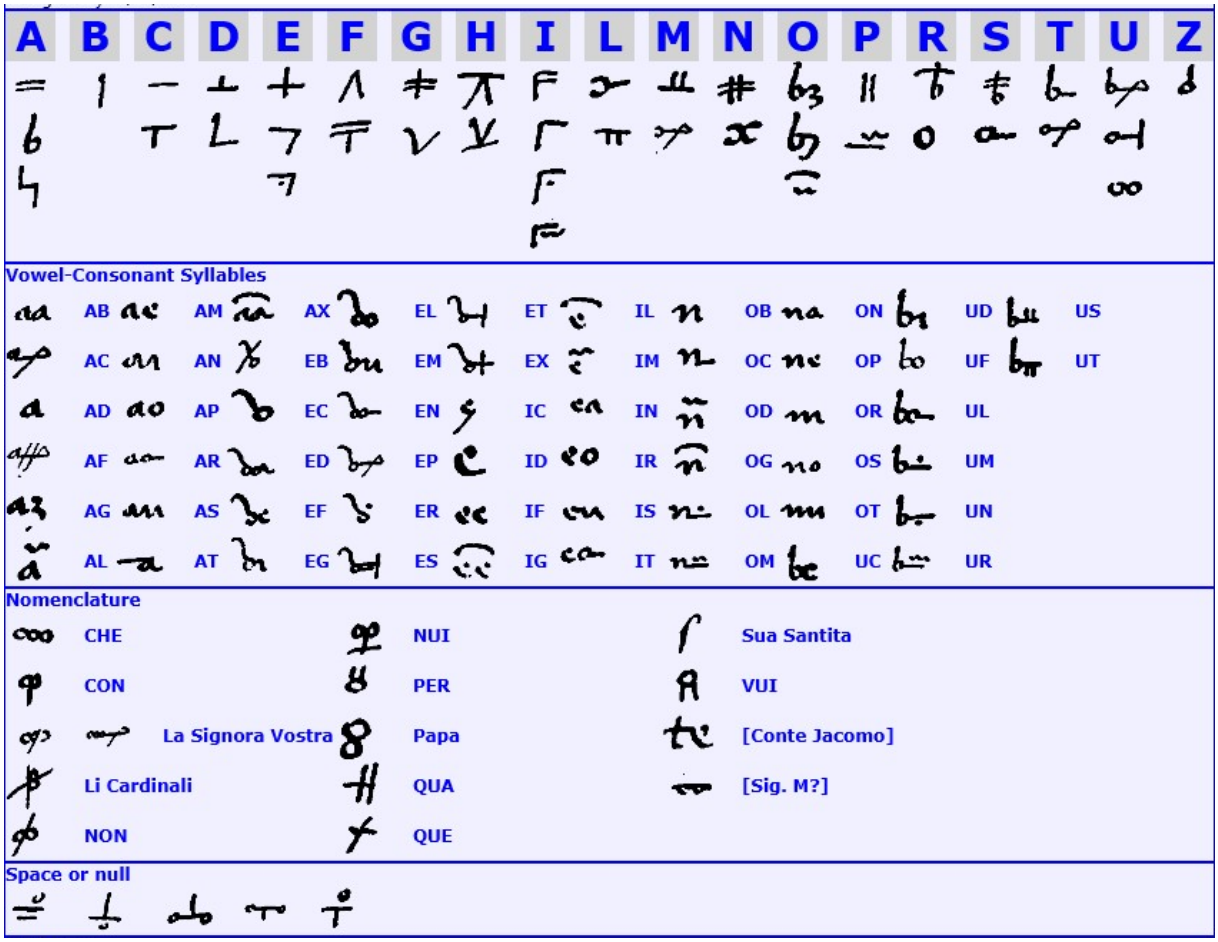
George Lasry, 2021. *Deciphering a Letter to Louis XIV from his Ambassador to the Dutch Republic, le Comte d'Avaux, 1684*. HistoCrypt 2021.

Cécile Pierrot, Camille Desenclos, Pierrick Gaudry, and Paul Zimmermann, 2023. *Deciphering Charles Quint (A diplomatic letter from 1547)*. HistoCrypt 2023.

Marcello Simonetta, 2023. *Svelati i segreti delle lettere di Castiglione alla vigilia del Sacco di Roma* [Accessed: November 2023]. Storia in Rete. https://storiainrete.com/svelati-i-segreti-delle-lettere-di-castiglione-alla-vigilia-del-sacco-di-roma/

Satoshi Tomokiyo, 2023. *Cryptiana, Articles on Historical Cryptography* [Accessed: November 2023]. http://cryptiana.web.fc2.com

**Appendix 1 – The reconstructed key for the Ottone de Corretta cipher**



**Appendix 2 – The reconstructed key for the Brasset-Mazarin cipher**