Transcription of Historical Ciphers and Keys

Beáta Megyesi
Department of Linguistics and Philology
Uppsala University, Sweden
beata.megyesi@lingfil.uu.se

Abstract

Historical ciphertexts and keys contain a wide range of symbols from digits and letters from known alphabets to various types of graphic signs. To be able to study ciphertexts and keys empirically in large(r) scale, consistent representation of the symbol systems used in ciphers is inevitable. In this paper, we present guidelines for transcription of ciphertexts, keys and cipher-related cleartext documents. We hope that the guidelines contribute not only to the systematic and consistent text representation across ciphertexts and keys, but also help in more accurate and reliable transcriptions.

1 Introduction

Usually, the first necessary, albeit time-consuming and probably least fun step in attacking a hand-written cipher is the conversion of the cipher image into a machine-readable format. The goal is to represent the ciphertext image as a text file, allowing various types of analyses. The process of converting the ciphertext image into a text document is called transcription. And the first, often cumbersome, albeit fun step in this process is the identification of the symbols, also called glyphs, in the ciphertext. During transcription, we need to identify and uniquely represent each symbol type by investigating the glyphs and their context. For this purpose, we usually create a transcription scheme, where each symbol type has its own and unique text representation. Then, we transcribe each glyph in the ciphertext according to our transcription scheme. We type in all glyphs, symbol by symbol, as they appear in the ciphertext in the text file.

The ciphertext alphabet might contain a wide range of symbols, such as letters, digits, punctuation marks, or other graphic signs. The identification of the symbol set is often unproblematic if the ciphertext is built up of some standard symbol set(s), such as digits (0-9), the Roman alphabet (a-z, A-Z), or a combination of the two. These symbols can be typed in easily and fast on a keyboard, and saved as a text file using some character encoding, such as a Unicode (UTF-8) format. However, ciphertexts often include a palette of symbols from various alphabets (Roman and Greek), graphic signs (Zodiac symbols or alchemical signs), diacritics, and punctuation marks (dots, commas). Nice examples of ciphertexts with mixed symbol sets is the Borg[1] (Aldarrab, 2017) and the Copiale[2] (Knight et al., 2011) ciphers with available transcriptions stored in the DECODE database (Megyesi et al., 2019).

The identification of the cipher alphabet is far from easy as symbols might look similar to each other although they represent different plaintext entities. Symbols can have diacritics, dots or other marks attached to them, or these can be unintentional ink spots or dirt that should not be part of the transcription. While the encoded sequences in ciphertexts are usually meticulously written and often segmented glyph by glyph to avoid any kind of ambiguity for the receiver to be able to decode the content, sequences of connected symbols or sloppy handwriting are also frequent. In addition, the ciphertext might be embedded in cleartext, i.e. texts written in a known natural language.

Presumably, the transcriber strives for a simple and fast transcription process and chooses a mnemonic, easy to remember transcription scheme. He/she makes decisions about how to represent each symbol type, and

[1] https://cl.lingfil.uu.se/~bea/borg/
[2] https://cl.lingfil.uu.se/~bea/copiale/

how to transcribe each glyph, space, punctuation mark, along with margin notes, catchwords, and cleartext sequences. While the transcriber freely designs his/her transcription principles, we get a large variety of transcriptions which makes it hard to comparatively study these historical sources.

The aim of this paper is to present transcription guidelines to represent ciphertexts and keys with a great variation of symbol system in a text format. First, we give an overview of the basic principles for transcription, then we describe the guidelines for the transcription of ciphertext images and keys, followed by cleartext images representing the original plaintext or a text related to the ciphertext, for example in a letter correspondence. Lastly, we conclude the paper.

2 Transcription of Ciphers

Transcription is the systematic representation of language in written form, an effort "to report—insofar as typography allows—precisely what the textual inscription of a manuscript consists of" (Meulen and Tanselle, 1999). In what follows, we apply the terminology concerning writing systems as defined by Sproat (2006).

Not surprisingly, there is no standard convention for the transcription of manuscripts due to the great variety and heterogeneous nature of historical written sources (Meulen and Tanselle, 1999). Transcription is always based on the transcriber's interpretation, and can be said to be non-neutral given that the transcriber needs to decide upon how detailed or close the transcription should be to the original image (Rosenberg, 2006). Various considerations can be taken to decide which reading is the most likely to the original, and how detailed the transcription shall be. Such details can include the distinction of letters (e.g. i with or without a dot), capitalization and graphic emphasis such as section titles, abbreviations in original and their expansions, gaps and damages, as well as the scribe's self-corrections, in particular insertions, replacements and changes (Cipolla, 2018).

The level of the detail required depends on the aim (Koester, 2010). Even in a single manuscript written by one scribe, the shape of the letters can vary greatly, and deletions, additions, notes, marks can occur in many different ways which influence our interpretation (Driscoll and Pierazzo, 2016). Knowledge of the historical context, the culture and society in which the manuscript was produced is also relevant. A high level of granularity in the transcription provides insight into the practice of copying and its procedural character (Burnard et al., 2006) which might needed for editorial work for philologists and historians.

Our main purpose of transcription is to replicate the text content of the manuscripts to create a machine-readable text file for (crypt)analysis. In the case of ciphers, being it ciphertexts, keys, plaintexts or cleartexts, the most important task is to map the symbols in the ciphertext onto symbol representation as a written language. Transcription is rather straigthforward if the symbol set of the cipher belongs to a known script, a writing system of a particular language. However, transcription is challenging when it comes to ciphers — while written language is an idealization, made up of a limited set of clearly distinct and discrete symbols (Piotrowski, 2012), ciphertexts are made up of symbols of a potentially unlimited number taken from various alphabets (e.g. Latin or Greek) and arbitrary symbol sets (e.g. Zodiac or alchemical signs).

The transcription conventions we apply need to be easy-to-use (Kline and Perdue, 2020) and to put into practice, albeit precise to be useful for decryption purposes. The transcription shall be i) computer-readable, ii) stored as plaintext files, iii) in a uniform encoding allowing to represent various scripts and symbols. All symbols that are part of the cipher shall be present in the transcription and represented so that all necessary information that might have impact on the interpretation and decryption of the manuscript is present. The transcription shall reflect the intention of the encoder and remain as faithful to the original manuscript as possible, which includes retaining the original line length, capitalization, punctuation or lack thereof, spelling and misspellings, additions, and marks.

In addition, information about the transcription shall be provided in terms of meta-

data containing information about the original image(s) of the encrypted manuscript (Desenclos, 2016), and the transcription process with possibility to leave comments. Metadata should follow the TEI guidelines (TEI Consortium, 2020) and as for the format, XML is recommended but the transcription process might become slow and time-consuming. We leave to the transcriber to decide upon his/her own metadata and in the following, we give only a minimal set to serve as suggestion, as an example. For our current purposes in the DECRYPT project, we store metadata about the encrypted source directly in the DECODE database, and we do not need a repeated set of metadata in the transcription files. Here, we store information about the type of the encrypted source (ciphertext, key, cleartext), the name of the folder and the image where the original is located, and the name or ID of the transcriber. We also store information about the transcription, the date when the transcription was created, and the approximate time it took to transcribe the image along with the transcription method so we can compare various methods. Examples are manual transcription by typing or dictating, or semi- or fully automatic methods using hand-written text recognition. The transcriber can also leave comments about difficulties and problems.

The transcription guidelines presented in this paper constitute a summary of a detailed set of guidelines for encrypted sources, presented in (Megyesi, 2020) with many illustrations and examples. The guidelines have been applied to the transcription of several hundred of encrypted manuscripts and stored in the DECODE database (Megyesi et al., 2019). The transcriptions we create serve for the decryption and analysis of ciphers, including ciphertexts, keys, and cipher-related cleartext documents. The guidelines are continuously developed as we stumble on new types of encrypted sources. In the following, we describe the typical problems and cases and describe how we deal with them.

3 Transcription of Ciphertext

Ciphertexts contain symbol sequences, letters from existing alphabets, digits, other graphic signs, or a mixture of these. Ciphertexts might contain spaces, or the symbols follow each other one by one without any space or other marks between words, so called scriptura continua, used to hide word boundaries. Similar to historical text, punctuation marks are not frequent, sentence boundaries are typically not marked, and capitalized initial characters in the beginning of the sentence are usually missing, but they might appear. On the other hand, dots, commas or other marks might be used to indicate special codes or code groups. We can also find nulls in ciphertexts, i.e. symbols without any corresponding plaintext characters to confuse the cryptanalyst to make decryption even harder.

3.1 Metadata

Each transcript file of a particular cipher (which may consist of multiple images) starts with metadata with information about the file. Each line is initiated by '#' followed by a transcription attribute and its value as illustrated in Figure 1.

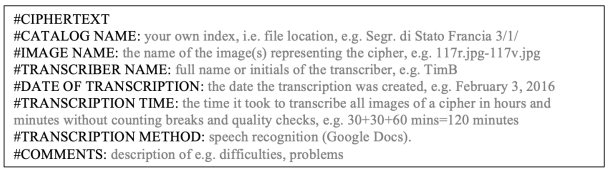

#CIPHERTEXT
#CATALOG NAME: your own index, i.e. file location, e.g. Segr. di Stato Francia 3/1/
#IMAGE NAME: the name of the image(s) representing the cipher, e.g. 117r.jpg-117v.jpg
#TRANSCRIBER NAME: full name or initials of the transcriber, e.g. TimB
#DATE OF TRANSCRIPTION: the date the transcription was created, e.g. February 3, 2016
#TRANSCRIPTION TIME: the time it took to transcribe all images of a cipher in hours and minutes without counting breaks and quality checks, e.g. 30+30+60 mins=120 minutes
#TRANSCRIPTION METHOD: speech recognition (Google Docs).
#COMMENTS: description of e.g. difficulties, problems

Figure 1: Metadata of the ciphertext.

3.2 Content

Next, the content of the page is transcribed. Each new image in a cipher starts with a new comment line with information about the name of the image followed by a possible comment line, see Figure 2. Then, the actual content of the ciphertext is transcribed.

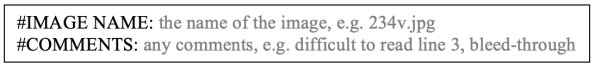

#IMAGE NAME: the name of the image, e.g. 234v.jpg
#COMMENTS: any comments, e.g. difficult to read line 3, bleed-through

Figure 2: Metadata of one page ciphertext.

The transcription is carried out symbol by symbol and row by row. This means that numbers are transcribed as numerals in ASCII, as typed in on the keyboard. The same applies to the letters in the Latin alphabet including capitalized letters, as well as punctuation marks. For other symbols, we use the Unicode name

representation where the name of the symbol is given following the Unicode standard.

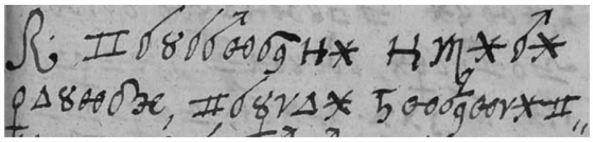
Handwriting varies greatly not only between individuals but also for the same writer, which is why transcription of ciphertexts containing special symbols is especially challenging.

The transcription shall represent the original ciphertext shown in the image, keeping line breaks, spaces, punctuation marks, dots, underlined symbols, and cleartext words, phrases, sentences, paragraphs, as shown in the original image.

3.2.1 Line breaks, Spaces, Punctuation and Diacritical Marks

Line breaks are kept so that when a new line starts, a new line is added in the transcription.

Space (' ') is represented as <SPACE> if it is clear from the ciphertext that space might indicate word boundaries, i.e. appear on regular basis in every line in a systematic way, as illustrated in Figure 3.

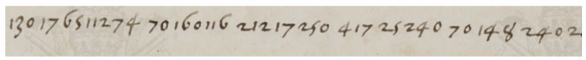


R **<SPACE>** gemini conjunction taurus conjunction ironore arsenic cancer H sextile
<SPACE> H virgo sextile ironore sextile
copperore fire taurus arsenic delta ? **<SPACE>** gemini conjunction taurus aries sextile
<SPACE> 5 arsenic cancer arsenic aries sextile gemini

Figure 3: Transcription of a cipher with graphic signs represented as Unicode names and word boundaries marked as <SPACE> in the ciphertext.

If space occurs, but apparently not in a systematic way, just happen to be there, the space can be transcribed with two or more space characters written in ASCII ' ' in the transcription, as illustrated in Figure 4. The reason for allowing several space characters is that a larger space in the original might mark word boundaries which the encryptor unintentionally left there when encrypting the manuscript, which can be helpful in the decryption process as they might denote word boundaries.

Punctuation marks such as periods, commas, and question marks are transcribed as such. Sometimes, punctuation marks (e.g.



130 176511274 70160116 21217250 417 2524 0 70 14 8 24 02

Figure 4: Transcription of a cipher with digits represented as ASCII characters and space marked as ' '.

dots, commas, accents, underscores) appear above or under specific symbols. It could be ink splash, but if they appear in a systematic way, they are transcribed as well. If the mark appears above the symbol, the sequence is transcribed as the symbol, followed by '^' and the specific mark (e.g. dot or comma). If the mark appears under the symbol, it is marked by an '_' placed between the symbol and the mark '.' (e.g. _.). Similarly, underlined symbols are marked with '_' (double underscore) immediately following the symbol, except when the whole ciphertext is underlined. Sub- and/or superscripts shall be indicated on all individual symbols in a sequence of symbols.

Example of some special symbols and their transcription is given in Figure 5. To avoid ambiguous cases for symbols with sub- and/or superscript, we mark the sub- and the superscript in brackets in the form *SYMBOL{superscript}{subscript}*.

	Glyph	Transcribed as
Dot on top		3^.
Accent on top		3^'
Dot on bottom		3_.
Dot on top and bottom		3{^.}{_.}

Figure 5: Transcription of symbols with diacritical marks.

3.2.2 Symbols

Symbols from other alphabets, such as Greek letters (α, β) Roman numerals (I, II), or graphic signs, such as the alchemical or Zodiac signs are also common in ciphertexts. To transcribe those, we use their Unicode representation transcribed by its Unicode name

which then can be automatically converted to Unicode code to visualize the symbol in some font. Figure 6 illustrates the Zodiac signs, each with its Unicode name and code, followed by the glyph.

If the symbol cannot be covered by the symbols from some common alphabet (Latin and Greek) or digit (Arabic or Roman), the transcriber should look at the Zodiac signs first, followed by the alchemical signs as those symbols occur often in (European) encrypted manuscripts. If it is not possible to find any similar symbol among them, a symbol that reminds the most of the original can be searched for in the large Unicode table of symbols. What is important to keep in mind, that the symbol is transcribed with a unique name to make it distinguishable from the other symbol types in the cipher.

Name	Code	Glyph
aries	2648	♈
taurus	2649	♉
gemini	264A	♊
cancer	264B	♋
leo	264C	♌
virgo	264D	♍
libra	264E	♎
scorpio	264F	♏
sagittarius	2650	♐
capricorn	2651	♑
aquarius	2652	♒
pisces	2653	♓

Figure 6: Zodiac signs.

An example of the transcription of a ciphertext with alphabetical characters (Roman and Greek) and graphic signs consisting of Zodiac and alchemical signs is shown in Figure 7 along with the transcription indicated by the Unicode symbol name, its automatic conversion to Unicode codes, and lastly the final visualization of the transcription.

Uncertain symbols are transcribed with added question mark '?' immediately following the uncertain symbol. Possible interpretations of a symbol can be transcribed using the delimiter '/'. For example, if it is not clear if a symbol represents a 0 or 6, it is transcribed as '0/6?'. It is highly desirable that all symbols are transcribed somehow, and no symbols are left out in the transcription for reliable decryption. The question mark ensures that all symbols have some representation in the transcription.

3.2.3 Catchwords

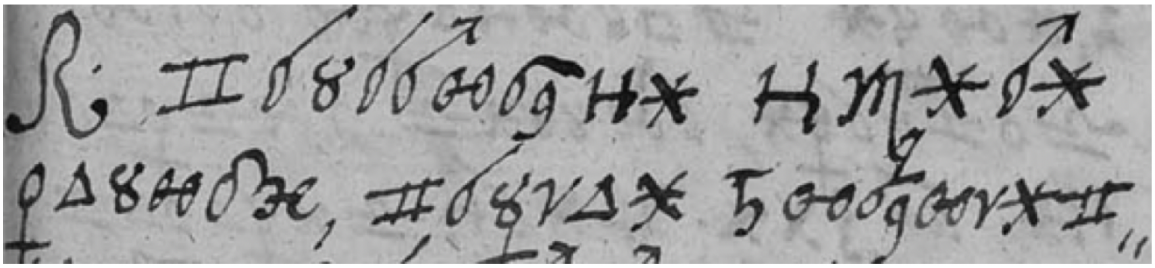
Historical manuscripts might contain catchwords placed at the foot of the page to mark page order (instead of digits), as illustrated in Figure 8. Catchwords are a sequence of symbols anticipated as the first symbol(s) of the following page. In ciphers, catchwords might denote an actual word, unintentionally, and transcribed as *<CATCHWORD Symbol_Sequence>*, as exemplified in Figure 8.

3.2.4 Notes in Margins

Sometimes ciphertexts are also included in the margins. This happens basically for two reasons: for corrections indicated in the ciphertext with a mark and the item is written in the margin, or the ciphertext continues in the margin to save space.

Transcription shall always reflect the intention of the encoder, i.e. the corrected segments as visualized in the original are transcribed. For example, if numbers are crossed-off in the original, these are not transcribed. If such cases occur, the transcriber leaves a comment about it in the comment line of the metadata. Similarly, insertions of corrections between symbols are transcribed, as they intended to appear. Ciphertext/cleartext written in the margin is added into the specific place as indicated by the given mark in the original. In Figure 9, the '+' written by the encoder intended to insert the cipher sequence written in the margin marked in red, and the transcription mirrors the intention of the encoder by directly adding the cipher sequence in the margin to the ciphertext.

Notes in the margin that are not corrections are transcribed after the transcription of the ciphertext, initially marked by a comment line with a short description that the upcoming sequence is a note in the left or right margin.



Transcription by Unicode name:
R gemini conjunction taurus conjunction ironore arsenic cancer H sextile <SPACE> H virgo sextile ironore sextile
copperore fire taurus arsenic delta ? <SPACE> gemini conjunction taurus aries sextile SPACE 5 arsenic cancer arsenic aries sextile gemini

*Reproduced by Unicode codes *automatically*:*
R <SPACE> 264A 260C 2649 260C 2642 29df 264b H 26b9 <SPACE> H 264d 26b9 2642 26b9 2640 25b3 2649 29df 03b4 ? <SPACE> 264a 260C 2649 2648 25b3 26b9 SPACE 5 29df 264b 29df 2648 26b9 264a

Final representation for visualization:
R <SPACE> ♊ ☌ ♉ ☌ ♂ ⊶ ♋ H ∗ <SPACE> H ♍ ∗ ♂ ∗
♀ Δ ♉ ⊶ δ x? <SPACE> ♊ ☌ ♉ ♈ Δ ∗ <SPACE> 5 ⊶ ♋ ⊶ ♈ ∗ ♊

Figure 7: Transcription of cipher with graphic signs and alphabetical characters, Zodiac signs marked in purple.

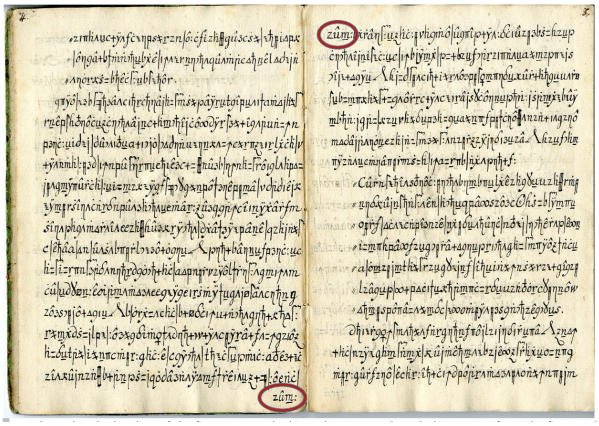


Figure 8: A cipher with catchword.

3.2.5 Ciphertext, Cleartext and Plaintext

The cipher sequences might be embedded in cleartext, i.e. non-encrypted text written in a natural language, or cleartext might be embedded in ciphertext. Cleartext embedded in ciphertext is illustrated in Figure 10 where the Spanish word sequence 'comè la mi cománda' is embedded in the surrounding ciphertext.

To be able to distinguish between ciphertext and cleartext sequences, the latter is clearly marked in brackets as *<CLEARTEXT LANG Letter/Word_sequence>* where the tag <CLEARTEXT... > denotes where the cleartext starts and ends as illustrated in the transcription in Figure 10. If the manuscript contains several lines of cleartext, each new line is represented by a new <CLEARTEXT... > tag. *LANG* represents the language the cleartext is written in, marked by a language *ID* as defined by ISO 639-1 two-letter codes [3] for languages (e.g. ES for Spanish, FR for French).

If there is some doubt about the cleartext/plaintext language, the language *ID* shall be defined as *UN*, indicating an unidentified language. For those cases where the cleartext does not necessarily constitute a certain language, such as dates (17.02.1725), years (1872)

[3] https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

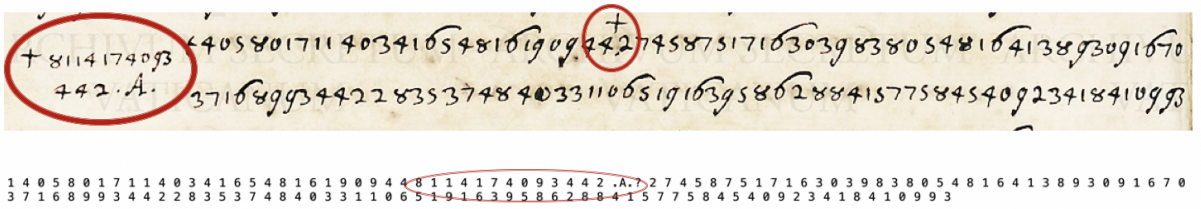


Figure 9: A ciphertext with corrections on the margin and its transcription.

```
130 176511274 70160116 21217250 417252407014824021013627012 27
220245845627670122721025024176 25 621224 0502484252617
1301222 2 <CLEARTEXT ES comè la mi coma^~da> 22225024701248474417 25242
507271216014424647238472525 6024472220295122462521 2
```

Figure 10: Transcription of a cleartext embedded in ciphertext.

or paragraph markers (P.25), the language tag N/A (not applicable) is applied, as shown in <CLEARTEXT N/A 1872>.

The cipher image might contain not only embedded (non-encrypted) cleartext, but also decrypted plaintext. We find decrypted plaintext written over the ciphertext sequences by the receiver, as illustrated in Figure 11. Similar to cleartext, plaintext is transcribed as *<PLAINTEXT LANG Letter/Word_sequence>* in a separate line.

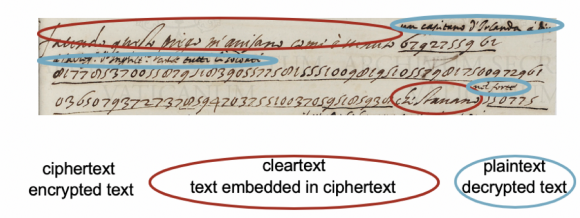


Figure 11: Cleartext and plaintext embedded in ciphertext.

3.2.6 Abbreviations

Sometimes we find abbreviations in the plaintext or cleartext sequences. Original text shall be transcribed as such, and in cases where abbreviations occur, the expansion of the abbreviation can be inserted after the abbreviated segment given as *<ABBR expanded-abbreviation>*. For example, *sre* in 'Del sre Bianco' is the abbreviation of signore and transcribed as in 'Del sre <ABBR signore> Bianco'.

4 Transcription of Keys

A key defines how each entity in the original plaintext shall be encrypted. Keys might contain substitution of not only characters in the plaintext alphabet, but also space to hide word boundaries, or nomenclatures where bigrams, trigrams, syllables, morphemes, common words, and/or named entities, typically referring to persons, geographic areas, or dates, are substituted with certain symbol(s). Punctuation marks or capital letters might occur in keys. A key might also contain nulls, i.e. symbols without any corresponding plaintext characters to confuse the cryptanalyst to make decryption even harder, explained in cleartext, or given as cipher symbol (Megyesi et al., 2019). Codes might also be present without any plaintext, serving as placeholders (Tudor et al., 2020).

The codes in a key might be of variable length. Each type of entity to be encrypted can be encoded by one symbol only, two symbols, three symbols, and so on. For example, the plaintext alphabet characters might be encrypted with codes using two-digit numbers, the nomenclatures with three-digit numbers, space with one-digit numbers, and the nulls with two-digit numbers, etc. Figure 12 illustrates a key based on homophonic substitution with nomenclature from the second half of the 17th century. Each sign in the alphabet is represented by at least one ciphertext symbol (e.g. A->18, m; B->20; C->19). The vowels and double consonants are assigned an additional ciphertext sign. The key also con-

tains encoded syllables with two-digit numbers or bigram characters (e.g. ba->65; be->66), followed by a nomenclature in the form of a list of Spanish words encoded with three-digit numbers or symbols (e.g. apustamiento->106). Keys might also include cleartext with explanation to (some parts of) the key. Similar to ciphertexts, metadata of the key is defined first, followed by the transcription and possible cleartext appearing in the original key.

4.1 Metadata

Before the actual transcription, original keys are described by a set of metadata, related to the transcription and the description of the key, each initialized by a hashtag (#) as defined in Figure 13.

4.2 Codes

After the metadata, the actual transcription of the content of the keys follows. The transcription guidelines for keys are partly based on the master thesis of Tudor (2019) and the transcription guidelines for ciphers (Megyesi, 2020). For keys, the same principles apply as for ciphertexts, when it comes to symbols described in Section 3.2.2, and cleartext sequences, which often contains explanations about the cipher key and explained in Section 3.2.5.

Since keys can be structured in many different ways, often as tables with or without explanations in cleartext, the graphical structure of the keys cannot be represented in any simple way in the transcription. Here, we make an interpretation of the content of the coding scheme instead. We list the key items as <*CODE-PLAINTEXT*> pairs where each unique pair is written in a line, first the code followed by the separator '–', then the plaintext unit, being it a character in the alphabet, syllable, word, null, or punctuation mark. Nulls are transcribed as <NULL> and missing plaintext of a code is transcribed as <EMPTY> (Tudor et al., 2020).

To illustrate the key representation, as shown in the key in Figure 12, the first three letters *A*, *B*, and *C* with their first code, are represented in the transcription as follows:

```
18 - A
20 - B
19 - C
```

A plaintext unit can be coded by several ciphertext symbols, such as in homophonic ciphers. In those cases, the possible codes are transcribed sequentially separated by a bar 'l' followed by '–' and the plaintext unit. For example, in our example in Figure 12, *A* can be coded not one but two possible ways, with the number *18* and the letter *m*. The alternative codes are transcribed in one line even when these are written in two lines in the original key, as illustrated below:

```
18 | m - A
20 - B
19 - C
```

Similarly, in case of polyphonic cipher keys where a ciphertext symbol can be mapped to several plaintext units, each plaintext symbol is listed with the code, separated by a bar ('l') in one line, no matter if they appear on separate lines in the original. For example, if the code *0* in the key might encode two plaintext letters, e.g. *a* and *t*, we would transcribe it as:

```
0 - a|t.
```

Please note that the separator bar '|' aimed for separating code or plaintext alternatives in keys is written in ASCII. However, if the ciphertext symbol represents the glyph '|' in the code itself, it is transcribed with its Unicode name 'verticalline'.

5 Transcription of Cleartext

Cleartexts are defined as non-encrypted plaintexts. These could be letters without any ciphertext that appear in the context of a cipher, e.g. in a letter correspondence, or it could appear embedded in ciphertext, as described in Section 3.2.5.

5.1 Metadata

The metadata for cleartext documents contains the information shown in Figure 14.

5.2 Cleartext Content

Next, the content of the image is transcribed. Each new image starts with a new comment line with information about the name of the image followed by a possible comment line, similar to Figure 2.

The transcription shall represent the original text shown in the image, keeping line

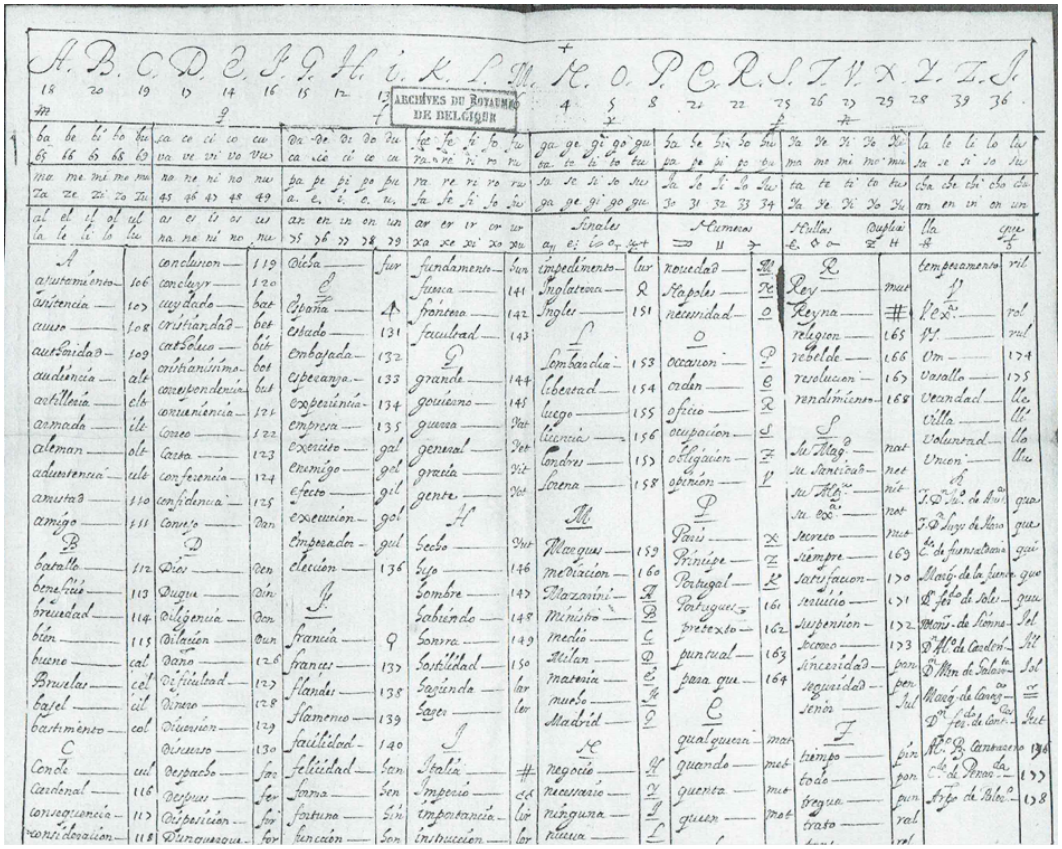


Figure 12: A key from the second half of the 17th century.

#KEY: original (keys that are reconstructed through cryptanalysis as opposed to original keys found in the archives are marked as reconstructed)
#CATALOG NAME: your own index, i.e. file location, e.g. ASV_ARM_XLIV_7-1
#IMAGE NAME: the name of the image(s) representing the key, e.g. 1287.jpg
#TRANSCRIBER NAME: full name or initials of the transcriber, e.g. CT
#DATE OF TRANSCRIPTION: the date the transcription was created, e.g. 22.04.2019
#TRANSCRIPTION TIME: the time it took to transcribe all images of the key in hours and minutes without counting breaks and quality checks, e.g. 1h 30min
#LANGUAGE: The language that is identified from the plaintext and cleartext present in the key according to the ISO 639-1 standard, e.g. FR for French, ES for Spanish, and UN for identified language, e.g. IT
#STATUS: [complete or partial]. Partial is used if part of the original key is known only.

Figure 13: Metadata of a key.

#CLEARTEXT
#CATALOG NAME: your own index, i.e. file location, e.g. Segr. di Stato Francia 3/1/
#IMAGE NAME: the name of the image(s) representing the cipher, e.g. 117r.jpg-117v.jpg
#TRANSCRIBER NAME: full name or initials of the transcriber, e.g. TimB
#DATE OF TRANSCRIPTION: the date the transcription was created, e.g. February 3, 2016
#TRANSCRIPTION TIME: the time it took to transcribe all images of a cipher in hours and minutes without counting breaks and quality checks, e.g. 30+30+60 mins=120 minutes
#TRANSCRIPTION METHOD: speech recognition (Google Docs).
#LANGUAGE: The language that is identified from the cleartext according to the ISO 639-1 standard, e.g. FR for French, ES for Spanish, and UN for identified language, e.g. IT
#COMMENTS: description of e.g. difficulties, problems

Figure 14: Metadata of a cleartext document.

breaks, spaces, punctuation marks, dots, underlined symbols, and cleartext words, phrases, sentences, and paragraphs, as shown in the original image. More specifically:

- Line breaks are kept so that when a new line starts, a new line is added in the transcription.

- Space is represented as space. Punctuation marks, such as periods, commas, and question marks are transcribed as such.

- Uncertain words or characters are transcribed with added question mark '?' immediately following the uncertain sequence. Possible interpretations of a symbol can be transcribed using the delimiter '/'. For example, if it is not clear if the word should be transcribed as *and* or *und*, all interpretations shall be transcribed with a question mark, as in 'and/und?'.

- Unidentified letters or words shall be marked with an asterix (*).

- Abbreviations. Original text shall be transcribed as such, and in cases where abbreviations occur, the expansion of the abbreviation can be inserted after the abbreviated word given as <ABBR expanded-abbreviation>.

6 Conclusion

We presented guidelines for a systematic and consistent transcription of historical encrypted sources: ciphertexts, keys, and cleartexts. Consistent transcription across ciphers provides the possibility to study and compare historical sources systematically in large scale. The guidelines might be also a useful resource in case we employ several transcribers of the same document for more accurate transcription. Our hope is that the guidelines will serve in getting a more accurate, unambiguous and consistent transcription within and across ciphertexts and keys, a first step taken to a standardized transcription of historical encrypted sources. Lastly, and most importantly, consistent transcription across symbols sets and scripts can also support (semi-)automatic transcription allowing sophisticated hand-written text recognition models — with or without human intervention — to take care of the tedious transcription process of historical manuscripts.

Acknowledgements

I would like to thank my colleagues in the DECRYPT project, in particular Michelle Waldispühl, George Lasry, and Nils Kopal for their valuable feedback on the guidelines. Crina Tudor deserves special thanks for her work on the transcription of historical cipher keys as part of her master thesis. Lastly, transcription of hundreds of ciphers would not have been possible without all students who take the time and effort to transcribe these fascinating historical sources. This work was supported by the Swedish Research Council, grant 2018-06074.

References

Nada Aldarrab. 2017. Decipherment of historical manuscripts. Master's thesis, University of Southern California. Master thesis in Computer Science.

Lou Burnard, Katherine O'Brien O'Keefe, and John Unsworth. 2006. *Electronic Textual Editing*. The Modern Language Association of America, New York.

Adele Cipolla. 2018. *Digital Philology: New Thoughts on Old Questions*. Libreria Universiteria Edizioni, Padova, Italy.

Camille Desenclos. 2016. Early Modern Correspondence: A New Challenge for Digital Editions. In *Digital Scholarly Editing: Theories and Practices*, UK. Open Book Publishers.

Matthew James Driscoll and Elena Pierazzo. 2016. *Digital Scholarly Editing: Theories and Practices*. Open Book Publishers, UK.

Mary-Jo Kline and Susan Holbrook Perdue. 2020. *A Guide to Documentary Editing*. The Association for Documentary Editing, Online edition, 3 edition.

Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2011. The Copiale Cipher. In *Invited talk at ACL Workshop on Building and Using Comparable Corpora (BUCC)*. Association for Computational Linguistics.

Almut Koester. 2010. Building Small Specialized Corpora. In *The Routledge Handbook of Corpus Linguistics*, pages 66–79.

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. 2020. Transcription of Historical Ciphers and Keys: Guidelines. `https://cl.lingfil.uu.se/~bea/publ/transcription-guidelines200221.pdf`. Version: February 10, 2020.

David L. Vander Meulen and G. Thomas Tanselle. 1999. A System of Manuscript Transcription. *Studies in Bibliography*, 52:201–212.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Morgan Claypool Publishers.

Robert Rosenberg. 2006. Documentary Editing. In *Electronic Textual Editing*, pages 92–104, New York. The Modern Language Association of America.

Richard Sproat. 2006. *A Computational Theory of Writing Systems*. Studies in Natural Language Processing. Cambridge University Press.

TEI P5 TEI Consortium. 2020. Guidelines for Electronic Text Encoding and Interchange (version 4.0.0). Accessed: 2020-04-27.

Crina Tudor, Beáta Megyesi, and Benedek Láng. 2020. Automatic Key Structure Extraction. In *Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt20*, Budapest, Hungary, June.

Crina Tudor. 2019. Studies of Cipher Keys from the 16th Century: Transcription, Systematisation and Analysis. Master thesis in Language Technology, Uppsala University, Sweden.