UPPSALA UNIVERSITET

Uncovering secrets from the past:

Research of historical cipher keys within the DECRYPT project

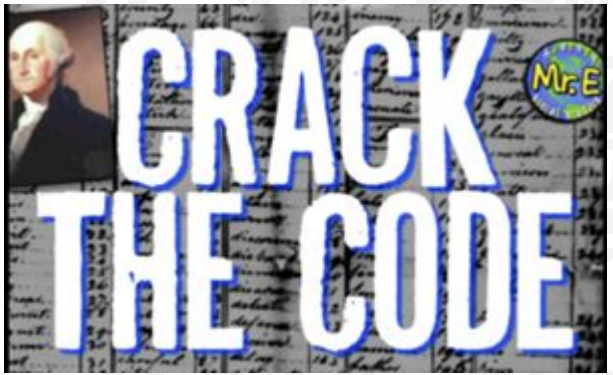
Crina Tudor
Department of Linguistics and Philology
Uppsala University

January 31st, 2022

Historical ciphers

- Thousands of ciphers are buried in archives
  - Diplomatic correspondence
  - Intelligent reports
  - Docs from secret societies
  - Private letters and diaries

•Not indexed, not marked up as such



- ✓ Need to bring to light the content for historical contextualization
- ✓ Require research infrastructure for historical cryptology
- ✓ Need of large-scale, systematic studies

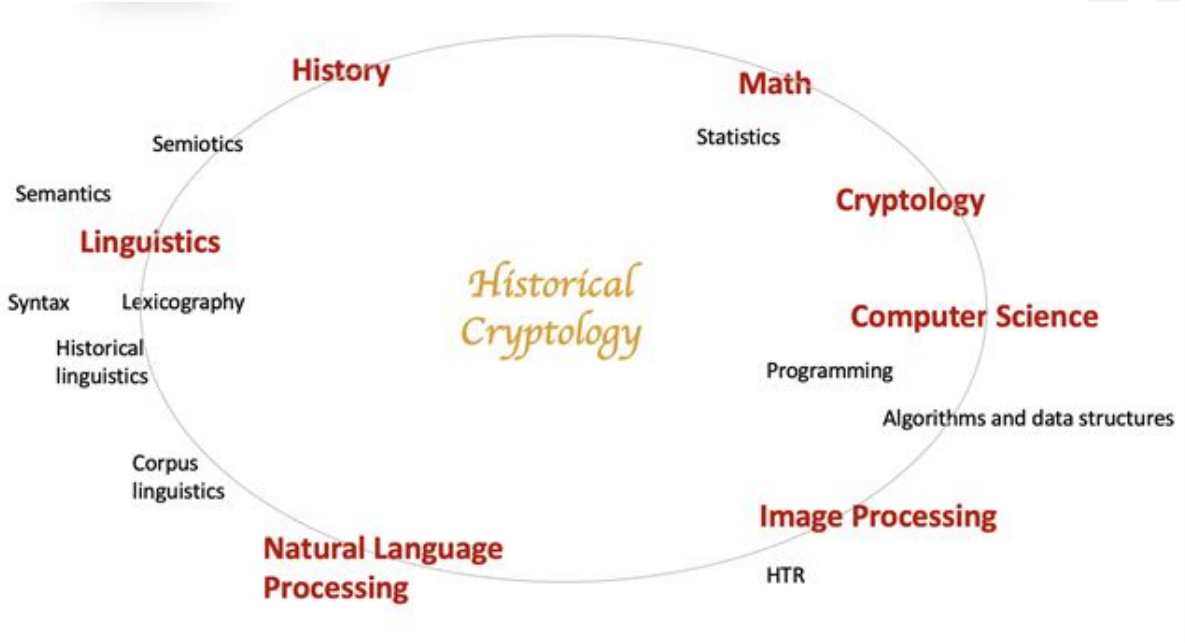
UPPSALA UNIVERSITET

The DECRYPT project

UPPSALA UNIVERSITET

Project aim

Establish a new cross-disciplinary scientific field of historical cryptology to decrypt and contextualize historical encrypted sources.



UPPSALA UNIVERSITET

The DECRYPT Portal

https://de-crypt.org/index.php

Resources:

- The DECODE database
- HistCorp

Tools:

- CryptTool2 - cipher breaking
- Transcript - interactive transcription tool
- Decoder - ciphertext-key mapping
- Anacode - ciphertext analysis
- Anakey - cipher key analysis

Anakey - purpose

- To provide a reliable transcription scheme for the transcription of historical keys
- To build a method for automatically identifying different types of keys
- To provide a statistical analysis for individual keys

UPPSALA UNIVERSITET

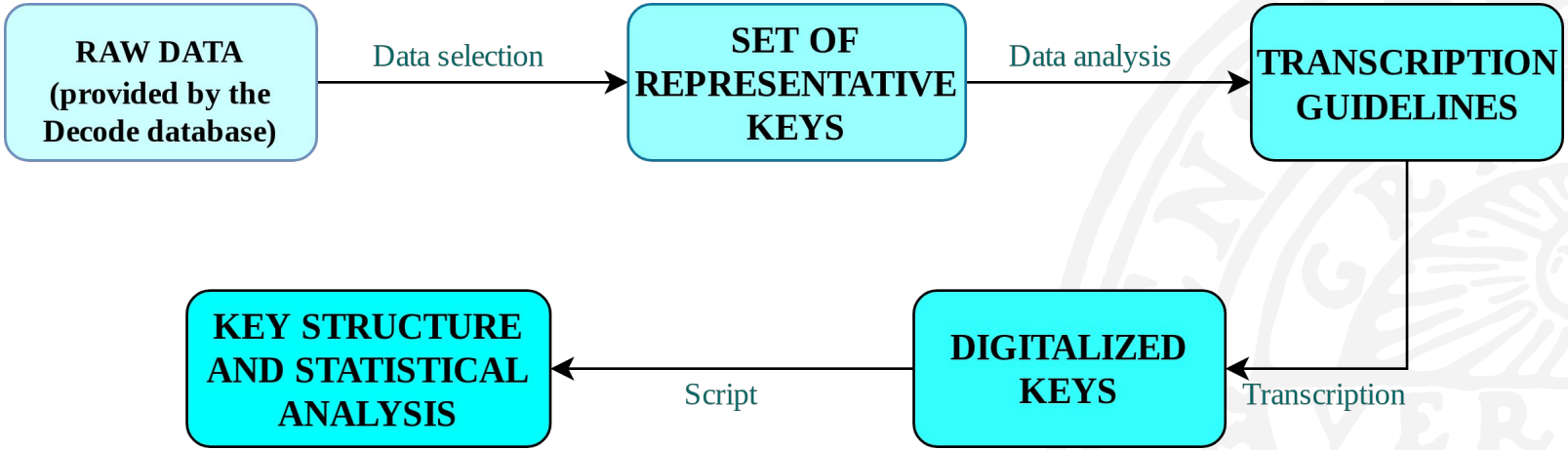
Background

- No previous study conducted on the structure of historical keys
- Most of the specialized literature focuses on ciphers rather than keys
- The few studies that discuss transcription methods for keys or ciphers tend to focus on individual instances and are not conducted on a large scale
- A completely automatic and accurate transcription is currently not achievable through OCR alone

UPPSALA
UNIVERSITET

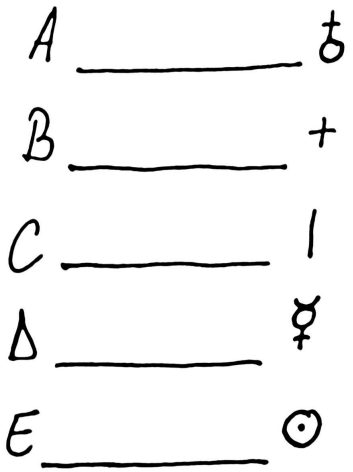
Key excerpt

Method


RAW DATA (provided by the Decode database)
Data selection
SET OF REPRESENTATIVE KEYS
Data analysis
TRANSCRIPTION GUIDELINES
Transcription
DIGITALIZED KEYS
Script
KEY STRUCTURE AND STATISTICAL ANALYSIS

Transcription Conventions

- Each transcription file is preceded by metadata
- We focus on methods for simple, homophonic and polyphonic substitution
- We focus on the non-ASCII characters
- We differentiate between 3 major symbol sets: Latin alphabet, digits, and graphic signs



Polyphonic substitution

Homophonic substitution

Key excerpt

UPPSALA UNIVERSITET

Key excerpt

#KEY: original
#CATALOG NAME: TNA_SP106/2_ElizabethI_f58(0069)
#IMAGE NAME: 3391.jpg
#LANGUAGE: FR EN
#TRANSCRIBER NAME: CT
#DATE OF TRANSCRIPTION: 10.04.2019
#TRANSCRIPTION TIME: 2h
#STATUS:complete
m - a
n - b
o - c
p - d
q - e
r - f
s - g
t - h
u - i
w - k
x - l
y - m
z - n
a - o
b - p
c - q
d - r
e - s
f - t
g - u
h - w
i - x
k - y
l - z

UPPSALA UNIVERSITET

Automatic Key Structure Extraction

- We build an automatic method for extracting key statistics
- Some of the information we can extract using our script includes:
  - type of symbols used for encryption
  - code structure (ngraphs)
  - unknown symbols
  - plaintext structure (ngrams)
  - code distribution

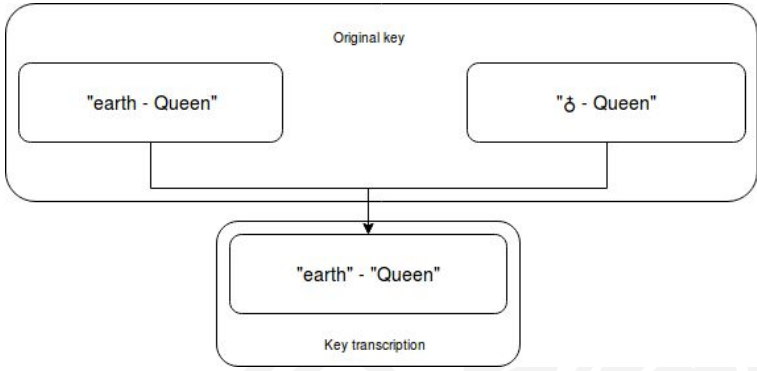
Error catching

Our script can catch the following types of errors:

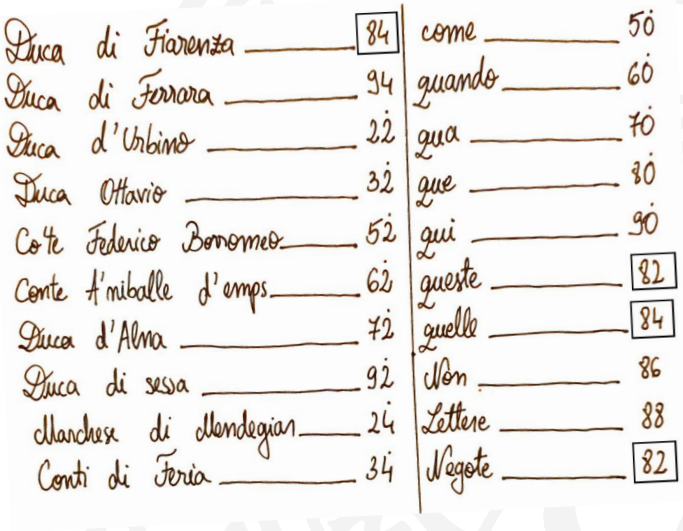
- metadata error
- delimitation error
- spacing error
- other

LANGUAGE: FR EN > #LANGUAGE: FR EN
89 a > 89 - a
89 - a > 89 - a

Future Work



- Our method provides a solid basis for further studies into the structure of keys
- We can still benefit from certain improvements:
    - eliminating the ambiguity when it comes to graphic signs
    - the way we handle sloppy keys
    - expanding our approach to other encryption methods



UPPSALA UNIVERSITET

Conclusions

- We provide a dependable transcription standard for historical keys
- We build a method for automatic key structure extraction
- We build a reliable basis for further large-scale comparative studies

Large-scale study

Megyesi B, Tudor C, Láng B, Lehofer A. *Key Design in the Early Modern Era in Europe*.
In: Proceedings of the 4th International Conference on Historical Cryptology (HistoCrypt 2021)

- *"At first, the substitution symbols were neither letters or numbers but fanciful signs like % or ·. But nobody has looked into when, in the later evolution, as nomenclators ran out of easily distinguishable symbols and began using numbers, the cipher secretaries began forming two-part nomenclators. This research requires merely examining the many nomenclators in the archives of Italy and France and timing and quantifying the change. I suppose it will be tough, living in Europe for a year and having an aperitif after a day examining antique manuscripts. But somebody should do it!"* (Kahn, 2008:58)

- This is exactly what we are up to….

UPPSALA UNIVERSITET

Research questions

- What types of keys were used in Europe between the 15th and 18th centuries? What were their specific characteristics?
- What was encoded and how?
- How did encryption evolve over time?
- Can we apply simple statistical methods to large-scale analysis of transcribed historical keys?

UPPSALA UNIVERSITET

Goal

- provide insight into the evolution of encryption
  - pilot: original keys from ca 1500-1800 in Europe
- provide a structural description of keys along with their morphological analysis and a typology
- investigate the key structure in terms of:
  - what is encoded - plaintext: languages, entity types
  - how is encoded - ciphertext: code types, symbols systems
- describe some trends of the code structure related to: time periods, geographic areas, ...

Method

- transcription guidelines for keys
  - common symbol representation across keys
- structural description of keys:
  - morphological (inner) structure
  - definition of key types
  - automatic structural description of keys
- match metadata info and structural description

UPPSALA UNIVERSITET

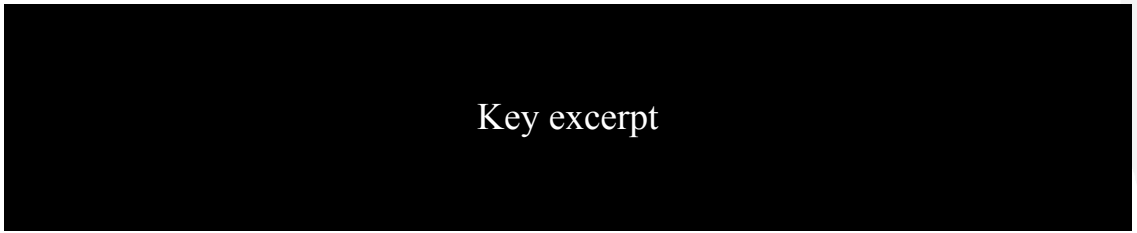
Data

- 1665 keys available in the DECODE collection from the 15th-18th centuries from Austria, Belgium, France, Germany, Hungary, Italy, Spain, the Netherlands, the UK, and the Vatican
- 26% of the keys have been manually transcribed

Automatic extraction of key structure

Extracted features from original key transcriptions:

- symbols types (digits, Latin, Greek, graphic signs, ...)
- code types:
  - no. of unique code types
  - unigraph, digraph, trigraph, 4+graph
  - fixed vs variable length
- plaintext types:
  - total number of unique plaintext units
  - unigrams, bigrams, trigrams, 4+grams



UPPSALA UNIVERSITET

Cipher type

Cipher type:

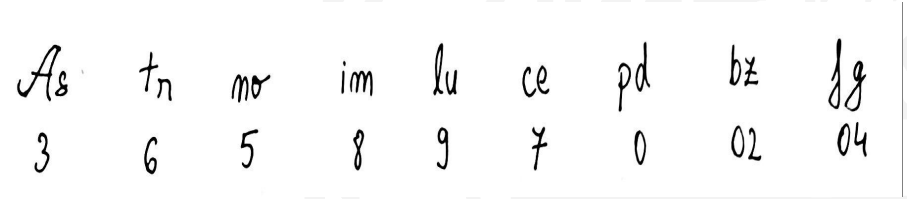
- defined on several levels:

  alphabet and nomenclature

- simple substitution
- homophonic substitution
- polyphonic substitution

Key excerpt

Key excerpt



Key excerpt

UPPSALA UNIVERSITET

Process

Key excerpt

```
#KEY: original
#CATALOG NAME: ASV_ARM_XLIV_7-1
#IMAGE NAME: 1287.jpg
#LANGUAGE: IT
#TRANSCRIBER NAME: CT
#DATE OF TRANSCRIPTION: 19.04.2019
#TRANSCRIPTION TIME: 2h
#STATUS:complete

#TYPE-NC: people, geographic names, common words

3 - A|s
6 - t|r
5 - n|o
8 - i|m
9 - l|u
7 - c|e
0 - p|d
02 - b|z
04 - f|g
00 - &con

22 - Il PP. N.S.,S.S.^ta
32 - Imp. S. M^ta Ces^a
42 - Re Filippo, Re Cath^co S. di^ta Cath^ca
52 - Re di Francia, Re dn^m s. di^ta Chr^ma
62 - Re di Portugallo
72 - Re di Polonia
82 - Re di Bohemia, S.ses^ta|queste|Anist
92 - Il Principe di Spagna
24 - Mons' di Vandomo
34 - Mad. la Regente, Regina di Francia
44 - Principessa di Portugallo
54 - Sig. Venetiani
64 - Sig. Suizze
74 - Duca di Savoia
84 - Duca di Fiarenza|quelle
94 - Duca di Ferrara
22^. - Duca d'Urbino
32^. - Duca Ottavio
52^. - Co'te Federico Borromeo
```

```
Metadata
#CATALOG NAME: ASV_ARM_XLIV_7-1
#LANGUAGE: IT
#STATUS:complete
#TRANSCRIPTION TIME: 2h
#TYPE-NC: people, geographic names, common words


Cipher symbols:digits, graphic signs

Total number of unique ciphertext symbols:78
        unigraphs:12
                out of which digits:7
        digraphs:66
                out of which digits:66
        trigraphs:0
        4+graphs:0


Total number of ciphertext symbols matched:78
No new ciphertext symbols were found.

Total number of unique plaintext units:94
        out of which unigrams:18
        out of which bigrams:0
        out of which trigrams:7
        out of which 4+grams:69
        out of which nulls:0
        out of which empty:0

Code distribution

Cipher type:polyphonic substitution

Code type:variable length

Number of codes encoding plaintext
        unigrams:9
        bigrams:0
        trigrams:6
        4+grams:64
        nulls:0
        empty:0

Distribution according to plaintext type (ciphertext:plaintext)
        1. Alphabet
        The alphabet has a uniform 1:1 distribution.
        2. Nomenclature
          1:1         65
          1:2         4
          1:3         1
          1:4+        0
```

UPPSALA UNIVERSITET

CSV output

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	Catalog name	Origin:city	Origin:country	Date	Language	Cipher symbols		Total # of unique ciphertext symbols	unigraphs	unigraph digits	digraphs	digraph digits	trigraphs	trigraph digits	4+graphs	4+graph digits		Total number of unique plaintext units	unigrams	bigrams	trigrams	4+grams	nulls	empty	cancellation		Cipher type	Code type
2	NAH_G15_CAPS_C_FASC_43_3	N/A	N/A	1703-01-01 - 1711-12-31	LA, EN	digits		298	0	0	58	58	238	238	1	1		254	71	93	8	127	6	0	0		homophonic substitution, polyphonic substitution, simple substitution	variable
3	NAH_G15_CAPS_C_FASC_43_2	N/A	N/A	1703-01-01 - 1711-12-31	LA	digits		301	9	9	90	90	201	201	0	0		284	33	111	16	134	7	0	0		homophonic substitution, simple substitution	variable
4	RAD_ARCH_Rakoczi_C64_4d2_25_3	N/A	N/A	1703-01-01 - 1711-12-31	LA	digits		94	9	9	75	75	10	10	0	0		45	96	1	1	17	4	0	0		homophonic substitution, polyphonic substitution, simple substitution	variable
5	RAD_ARCH_Rakoczi_C64_4d2_25_9	N/A	N/A	1703-01-01 - 1711-12-31	LA	digits		12	0	0	11	11	1	1	0	0		3	8	4	0	0	0	0	0		homophonic substitution	variable
6	TNA_SP106/4_Jamesl_ff19-20(0023-0024)	N/A	Spain	1614-01-01 - 1617-12-31	EN	digits		159	9	9	89	89	61	61	0	0		123	50	3	0	98	8	0	0		homophonic substitution	variable
7	RAD_ARCH_Rakoczi_C64_4d2_25_4	N/A	N/A	1703-01-01 - 1711-12-31	LA	digits		231	0	0	70	70	161	161	0	0		182	69	118	23	22	0	0	0		homophonic substitution	variable
8	Flemming in Vienna-Saxony 1761, HStAD, 10024, Loc. 08236/11, Bl. 24	Dresden	Saxony	1761-01-01 - 1761-12-31	FR	digits, Latin alphabet		100	21	1	74	72	5	5	0	0		100	24	47	4	24	1	0	0		homophonic substitution, simple substitution	variable
9	RAD_ARCH_Rakoczi_C64_4d2_25_5	N/A	N/A	1703-01-01 - 1711-12-31	LA	digits		266	0	0	66	66	200	200	0	0		217	78	85	6	100	0	0	0		homophonic substitution, simple substitution	variable
10	TNA_SP106/7_Anne_(0087-0089)-1	N/A	N/A	1702-03-08 - 1837-06-20	FR	digits		666	0	0	0	0	666	666	0	0		668	0	575	2	91	0	0	0		simple substitution	fixed
11	TNA_SP106/7_Anne_(0087-0089)-2	N/A	N/A	1702-03-08 - 1837-06-20	FR	digits		676	0	0	0	0	676	676	0	0		676	0	575	2	99	0	0	0		simple substitution	fixed
12	TNA_SP106/6_Charlesll_(0163-0164)	N/A	Belgium	1672-01-01 - 1672-12-31	LA	digits		356	7	7	90	90	199	174	58	0		293	2	56	102	176	20	0	0		homophonic substitution	variable
13	TNA_SP106/6_Charlesll_(0023-0024)	N/A	N/A	1660-05-29 - 1685-02-06	EN	digits		437	0	0	56	56	381	381	0	0		418	43	109	19	266	1	0	0		homophonic substitution, simple substitution	variable
14	RAD_ARCH_Rakoczi_C64_4d2_25_29	N/A	N/A	1703-01-01 - 1711-12-31	LA	digits		195	0	0	90	90	105	105	0	0		186	40	123	7	26	0	0	0		homophonic substitution, polyphonic substitution	variable
15	TNA_SP106/5_Charlesl_(0008-0009)	N/A	Spain	1630-01-01 - 1650-12-31	EN	digits		431	9	9	88	88	334	334	0	0		382	58	172	4	192	17	0	0		homophonic substitution	variable
16	Saxony 1761, HStAD, 10024, Loc. 08236/11, f. 29-30	Dresden	Saxony	1761-01-01 - 1761-12-31	FR	digits		30	5	5	25	25	0	0	0	0		32	28	1	1	2	0	0	0		homophonic substitution, simple substitution	variable
17	Copenhagen-Saxony-Hamburg 1761, HStAD, 10024, Loc. 08236/11, f. 26-28	Dresden	Saxony	1761-01-01 - 1761-12-31	FR	digits		114	10	10	96	96	8	8	0	0		130	24	46	6	41	2	0	0		homophonic substitution, polyphonic substitution	variable
18	Calenberg-Saxony 1761, HStAD, 10024, Loc. 08236/11, Bl. 3	N-CITY: Dresden	OUNTRY: Sax	1761-01-01 - 1761-12-31	FR	digits, Latin alphabet		107	34	9	73	71	0	0	0	0		108	23	52	4	28	0	0	0		polyphonic substitution	variable
19																												
20																												
21																												
22																												
23																												
24																												
25																												
26																												
27																												
28																												

UPPSALA UNIVERSITET

Trends

Goal:

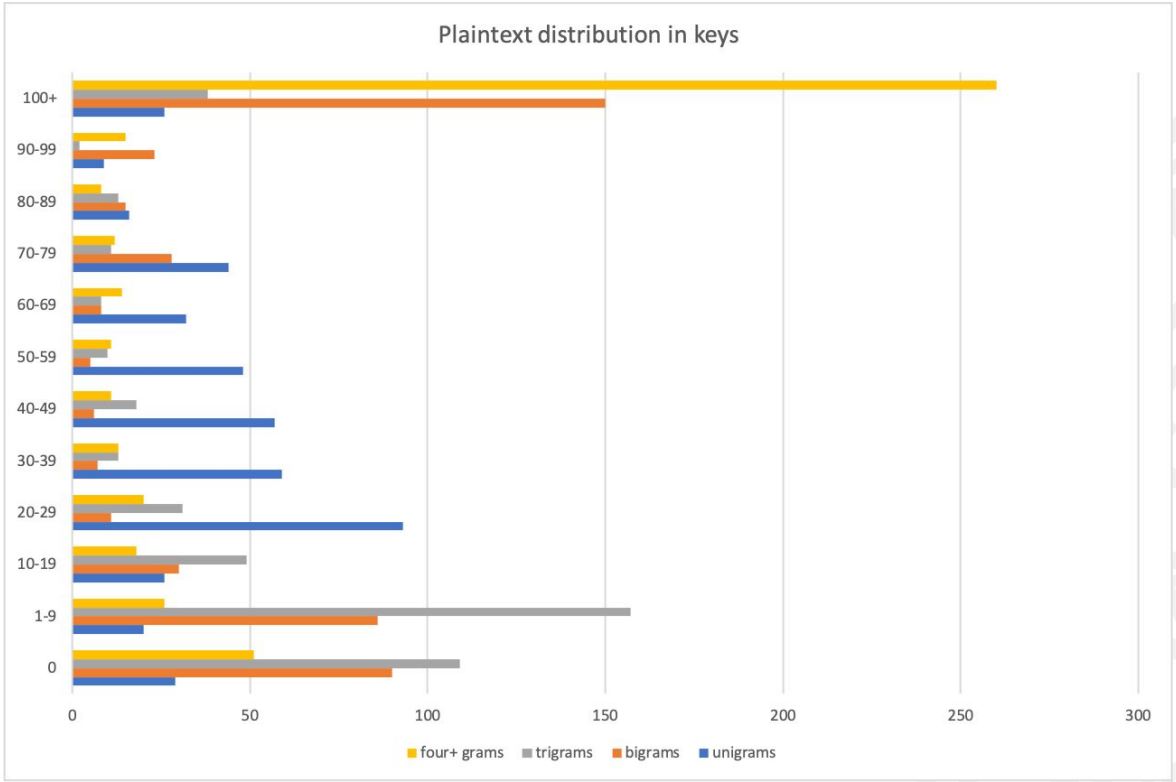
- investigate the trends throughout the 15th-18th centuries
- what have been chosen to be encoded and how

Data:

- 450 keys: automatically extracted from transcriptions
- 250 keys: manually extracted structural information without any transcriptions originating from the 15th and 16th centuries.
- 700 keys in total

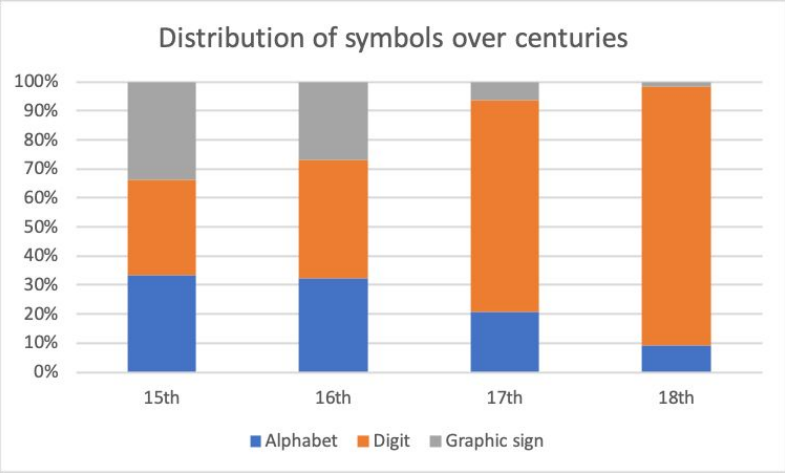
UPPSALA UNIVERSITET

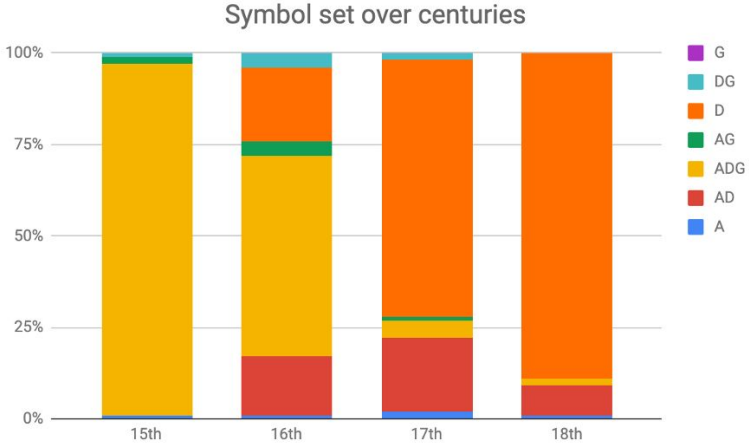
Plaintext



UPPSALA UNIVERSITET

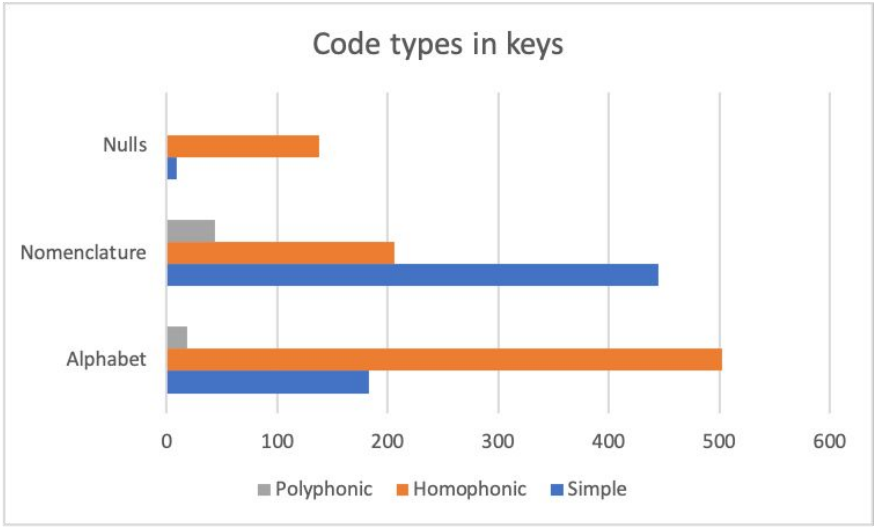
Symbol set

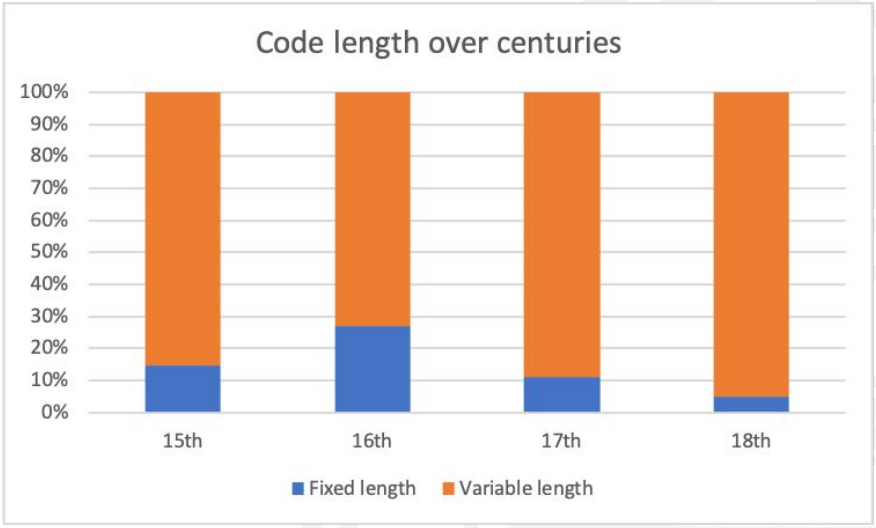




UPPSALA UNIVERSITET

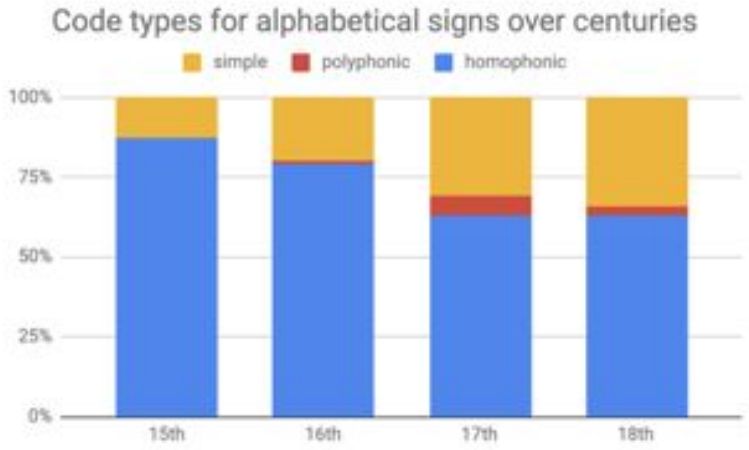
Codes

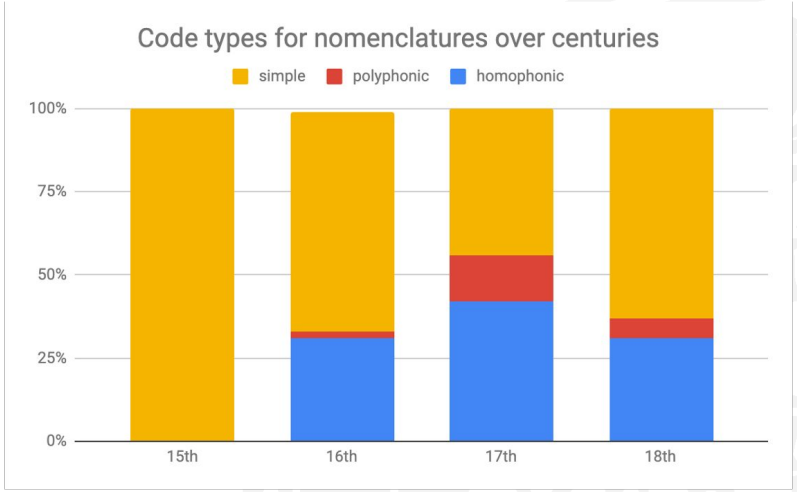




UPPSALA UNIVERSITET

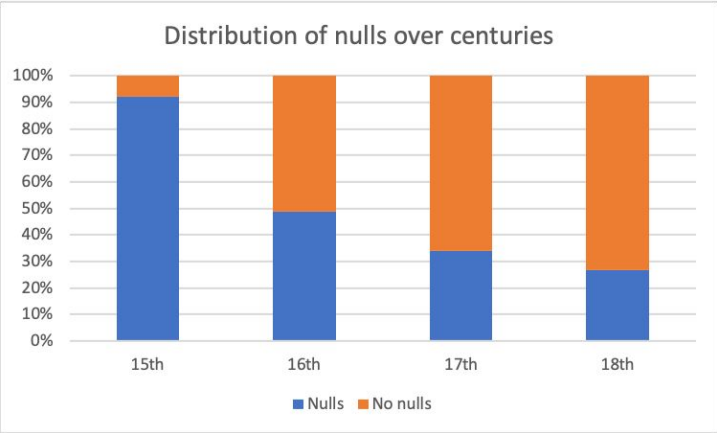
Codes





UPPSALA UNIVERSITET

Nulls



UPPSALA UNIVERSITET

Conclusion

- We investigated 700 cipher keys from the 15th to the 18th centuries
- We described the keys' internal structure and their morphology:
  - what have been chosen to be encoded and how
  - the type of the symbol set and the code structures used, and the changes and trends of each century.
- Keys evolved over time and their structure changed

UPPSALA UNIVERSITET

Conclusion

- Codes with various symbols including alphabets, digits, and graphic signs were dominating in the 15th century, digits only became more frequent and became the standard in the 18th century.
- The codes varied in length for alphabetical signs and nomenclatures throughout all centuries and codes with fixed length seemed to be most popular in the 16th century.
- Coding alphabetical signs was mostly homophonic, but simple substitution of letters became more frequent as the length of the nomenclatures increased over time.
- Nomenclatures were mostly encoded as simple substitution.
- Nulls have been frequently used in the 15th century and decreased significantly over time.
- Cancellation as phenomenon became ”popular” in the 18th century.

UPPSALA UNIVERSITET

Future/upcoming work

- include more data
- more precise metadata with location (GIS) (and person)!
  - correlations between features
  - add automatic key complexity (cont.)
  - JSON representation of automatic structural description
  - include the automatic structural description into the decrypt pipe ANAKEY

UPPSALA UNIVERSITET

Thank you!