UNIKASSEL VERSITÄT

Angewandte Informations Sicherheit

Insight from German WWI Codes

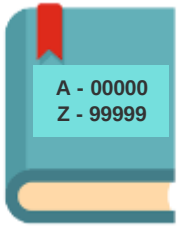
June 20, 2018
George Lasry
george.lasry@gmail.com

Agenda

- **German codes in WWI**
  - Overview
  - Mendelsohn's reconstruction of code 18470 (1918)
  - Insights from deciphering Genoa collection (2017)
  - The Stützel report (1917)
  - Room 40's Political Branch report (1918)
- **Solving codes - a hard problem**
  - Entropy and Unicity Distance of codes
  - A interesting algorithm
- **Unsolved code problems from WWI**

Types of Codes

- **One-part (ordered) code**
  - Words and codes in the same order (alphabetical)
  - Same physical book for encoding and decoding
- **Two-part (unordered) code**
  - Random order of codes
  - A.k.a. “hat” or “lottery” code
  - One book for encoding and another for decoding
- **Mixed code**
  - Base ordered code, with pages reshuffled
  - Words inside page partially reshuffled
- **Superencipherment**



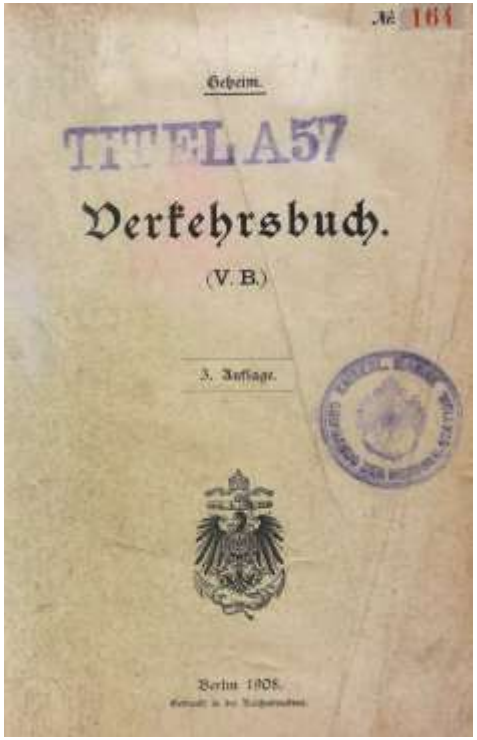
23030 ➔ 26530

German Codes in WWI

- **Kaiserliche Marine (German Navy)**
  - Mostly ordered codes with superencipherment
    - Signalbuch der Kaiserlichen Marine (SKM)
    - Flottenfunkspruchbuch (FFB)
    - Handelsverkehrsbuch (HVB)
    - Allgemeinefunkspruchbuch (AFB)
    - Verkehrsbuch (VB)
- **Deutsches Heer (German Army)**
  - Trench codes
  - Also ciphers (e.g. ADFGVX, FURGOD)

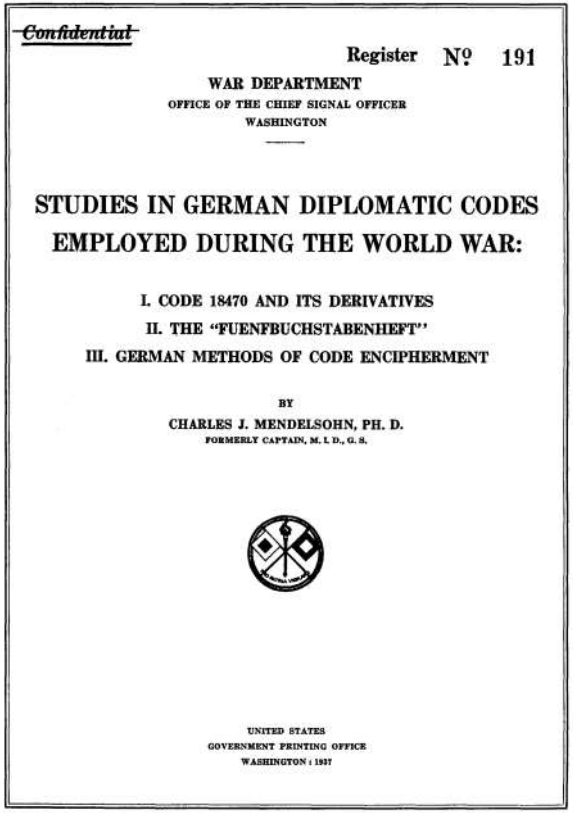
German Diplomatic Codes in WWI

- **Mixed codes**
  - New editions - reordering the pages
    - Order inside pages also changed – but in a regular manner
  - The 13040 family → 5950
  - The 18470 family → 3512, 1777, 2310, 12444, …
- **Hat codes**
  - 0042, 0053, 0064, 0075(7500), 0086, 0097
- **Naval attaché codes**
  - Verkehrsbuch (VB) with various superencipherments



The Mendelsohn Report

- **Partial reconstruction of 13040 by Room 40**
  - Provided to U.S. MI-8
  - Code 5950 – reordering of 13040
- **Mendelsonh report (1919, 1937)**
  - Describes in detail the structure of the 13040 and 18470, and how the 18470 was reconstructed
- **Hundreds of intercepted messages**
  - Mostly between Madrid and Berlin
  - Indicators 12444, 1777, 2310, 18470
  - Mendelsohn assumes those codes are related
  - Also assumes that the structure and reordering principles are similar to those of 13040/5950



~~Confidential~~

Register Nº 191

WAR DEPARTMENT
OFFICE OF THE CHIEF SIGNAL OFFICER
WASHINGTON

STUDIES IN GERMAN DIPLOMATIC CODES EMPLOYED DURING THE WORLD WAR:

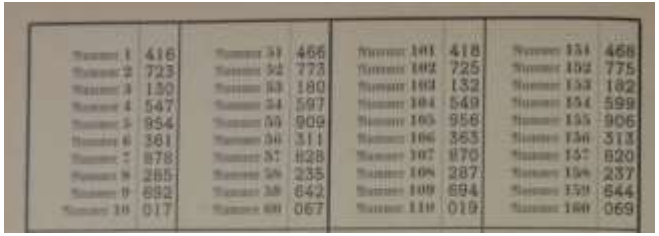
I. CODE 18470 AND ITS DERIVATIVES
II. THE "FUENFBUCHSTABENHEFT"
III. GERMAN METHODS OF CODE ENCIPHERMENT

BY
CHARLES J. MENDELSOHN, PH. D.
FORMERLY CAPTAIN, M. I. D., G. S.

UNITED STATES
GOVERNMENT PRINTING OFFICE
WASHINGTON : 1937

The Structure of the 13040 Code

- **Dreinummerheft (3-digit code)**
  - Numbers and dates
  - Used in other diplomatic codes
- **Vocabulary**
  - 189 pages with 100 items each
    - Blocks of 10 items in alphabetical order
    - Random ordering of blocks
- **Common phrases and prepositions**
  - Not always used
- **Grammatical instructions**
- **Places, persons and entities**



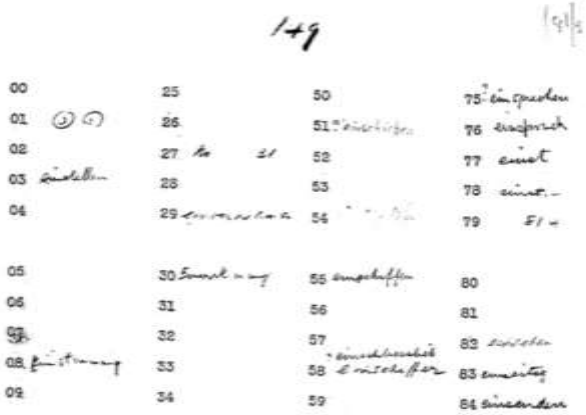


Figure 1. The top half of page 149 of a Room 40 copy of the 13040 codebook. Note that on each page groups for numbers (e.g. 14927 and 14979) and punctuation (e.g. 14901) are interleaved; decades are shuffled, but within decades the order is alphabetic.

The Mendelsohn Reconstruction of 18470 – Part 1

- **Step 1 – "Und"**
  - Frequent 5-digit code within 3-digit codes
  - Similarly – other common words
- **Step 2 – Messages in 13040 forwarded inside 124444**
  - Guesses for für, von/vom, Telegram Nummer
    - Before the 13040 parts
  - Assumes mapping of 12444, 1777, 2310, 18470 is similar to mapping between 13040 and 5950
    - Identification of several pages
    - Months (not in 3-digit code) – more pages identified

The Mendelsohn Reconstruction of 18470 – Part 2

- **Step 3 – Numerals not in 3-digit code**
  - Deterministic location in pages, as in 13040
  - Ordering of more pages
- **Step 4 – Extrapolating the page boundaries**
  - Based on expected number of words per page
  - Unsuccessful attempt using a German dictionary
  - Successful attempt using 13040 page boundaries
    - 13040 has 189 pages with words vs. 257 for 18470
  - Identifying words that are close in 13040

The Mendelsohn Reconstruction of 18470 – Part 3

- **Step 5 – Full mapping between 18470-12444-1777-2310 pages**
  - Pages of common ancestor “XX” – alphabetically ordered
- **Step 6 – Corresponding plaintexts-ciphertexts**
- **Step 7 – Additional progress for special cases**
  - Names of ships, places
- **The whole process take about one year**
  - Extensive set of index cards
    - All occurrences of codes
  - Interestingly, frequency count useful only for the most frequent codes
    - No use of bigrams or trigram frequencies

Insights from the Deciphering of the Genoa Collection - 2017

- **Computer software – only as an aid for book keeping**
  - Highly useful to implement mapping to related code
  - Also useful to remove superencipherment
- **Did not implement any cryptanalytic algorithm**
  - Only mechanical functions, e.g. lookup in book
- **Stuck in the process until the discovery of code 3512**
  - Implemented the mapping between 3512 and 18470 in software
- **Frequency analysis – useless**
  - Except for validating a few guesses

Top Discrete Words in the Genoa Decryptions (18470)

418 occurrences - und
405 occurrences - Nummer
401 occurrences - von
291 occurrences - der
290 occurrences - nach
281 occurrences - in
230 occurrences - für
227 occurrences - zu
209 occurrences - mit
196 occurrences - auf
193 occurrences - die
140 occurrences - aus
133 occurrences - an
111 occurrences - bitte
110 occurrences - vom
109 occurrences - Dampfer

Top words in regular text:
1. der/die/das
2. und
3. sein
4. in
5. ein
6. zu
7. haben
8. ich
9. werden
10. sie

Top Codes in the Genoa Collection (18470)

```
156 occurrences - 18654 in
115 occurrences - 18139 von
111 occurrences - 12462 und
109 occurrences - 10275 Dampfer [steamer]
106 occurrences - 02440 (Schluss der Depesche)
100 occurrences - 18470 (Chiffre 18470)
 95 occurrences - 05807 italienisch
 90 occurrences - 13788 bitte
 89 occurrences - 03818 Genua
 87 occurrences - 24922 hier
 83 occurrences - 35193 von Herff [the Consul]
 83 occurrences - 10057 mit
 81 occurrences - 01186 deutsch
 74 occurrences - 02441 (Schluss der Depesche)
```

Corpus size: 22995

Biased towards actual topics of interests and writing style

Top Code Bigrams in the Genoa Collection (18470)

```
36 occurrences - 27169 11516 für Größ. Generalstab
24 occurrences - 22366 30051 Auswärtig. Amt telegraphiert
21 occurrences - 15238 13293 goe ben
18 occurrences - 09852 53307 König Albert
17 occurrences - 05894 19586 Indisch Truppen
15 occurrences - 12192 12563 weitergebe/n zu
14 occurrences - 12563 12286 zu wollen
13 occurrences - 13788 15651 bitte gehorsamst
12 occurrences - 15651 04778 gehorsamst  anliegend
12 occurrences - 12149 12807 weide/n/t Mann
12 occurrences - 04778 30046 anliegend Tel.
12 occurrences - 19155 03818 Generalkonsulat  Genua
12 occurrences - 30854 11709 fünf hundert
11 occurrences - 35193 13788 von Herff bitte
11 occurrences - 30007 30854 tausend fünf
11 occurrences - 78675 53307 König Albert
```

Stronger bias – little match to general German

Top Code Trigrams in the Genoa Collection (18470)

```
14 occurrences - 12192 12563 12286 weitergebe/n zu wollen
12 occurrences - 13788 15651 04778 bitte gehorsamst anliegend
11 occurrences - 30007 30854 11709 tausend fünf hundert
11 occurrences - 30046 12192 12563 Tel. weitergebe/n zu
10 occurrences - 04778 30046 12192 anliegend Tel. weitergebe/n
10 occurrences - 18555 30007 30854 zwei tausend fünf
10 occurrences - 15651 04778 30046 gehorsamst anliegend Tel.
 9 occurrences - 07727 08007 22558 leu pol d
 9 occurrences - 27169 11516 24992 für Größ. Generalstab:(Alinea)
 9 occurrences - 18733 26224 03065 gleichlautend Botschaft Russland
 9 occurrences - 35193 13788 15651 von Herff bitte gehorsamst
 8 occurrences - 12648 28465 07606 kos mos Linie
 8 occurrences - 27169 11516 11582 für Größ. Generalstab:(Alinea)
 8 occurrences - 02077 14170 15238 Admiral kreuzer goe
 8 occurrences - 14170 15238 13293 kreuzer goe ben
 7 occurrences - 10275 78675 53307 Dampfer König Albert
```

The Stützel Report – 1917

Radio Section
O.H.L. (A.)
D. 468

Grand Headquarters of His Majesty 22 IX 1917

Result of Investigation of Cryptographic Systems used in Radio Traffic between Berlin and Madrid

The various types of radiograms observed here in traffic lp - ego, signed on the one hand by Zimmermann, Kuehlmann, Stumm, Bussche and on the other hand by Ratibor, Bassewitz, hence beyond doubt belonging to the telegraphic communications of the Foreign Office with the Imperial Ambassador Prince Ratibor in Madrid, can be divided essentially into two main groups:

18470 derivatives

1. Telegrams which have at the beginning the indicator groups 27082, 18470, 21894, 1777, 12444 with 4 and 5 digit groups up to 30900 and rare groups above 30900.

Hat codes

2. Cipher telegrams with the indicator groups 0053, 5003, 5300, 0000, 4343, 1357, chiefly with 4 digit groups.

The investigations were made on the basis of intercepted radiograms, i.e., with the same means which - at the very least - would be available to an unauthorized, hostile decipherer.

The Stützel Report – Hat Codes

To 2. Investigation of the telegrams with indicator groups 0053, 0000, 4343, 1357 yielded the following results:

a. the telegrams encoded with a code and enciphered by several methods. In 0000 encipherment is by simple transposition of the digits of the groups, in the other telegrams by addition and substitution according to frequently changing keys. (Supplement 1)

The discovery of the types of encipherment and thus the reduction of all these types to one basic type was carried out by one operator in 3 weeks.

The Stützel Report – Foreign Office Response

Your Honor's assertion that almost all cipher telegrams can be deciphered is untenable. If the matter were so easy, the German radio sections would probably not fail to decipher the Russian, English and French radiograms which they intercept. To my knowledge the German radio sections have only succeeded in partially deciphering the Italian radiograms; this may be

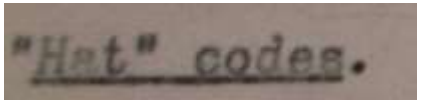
Representative of the Foreign Office
at Grand Headquarters Nr. 158, Secret.

General Headquarters, 23.III.17.

I bring to Your Honor's attention the foregoing letter of the Secretary of State of the Foreign Office to me. Please treat as confidential.

(Signed) Baron V. Lersner

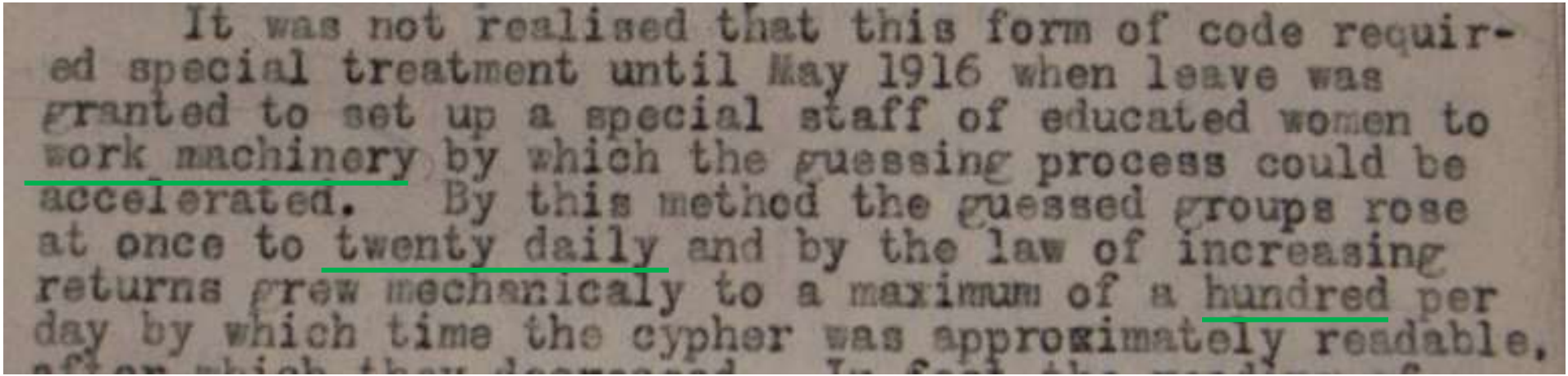
Room 40 – The Political Branch and Hat Codes


"Hat" codes.

solving another. Therefore the only means of reconstruction of such a code was a process of trial and error by which after comparing all the contexts in which a group occurred a guess could be made which in turn

might enable the same process to be repeated for an contiguous group. Given a mass of material an expert could in this way guess approximately four or five groups a day. As the life of such a cypher was only about eighteen months this form of cypher had always been considered practically insoluble.

Room 40 – The Political Branch and Hat Codes



It was not realised that this form of code required special treatment until May 1916 when leave was granted to set up a special staff of educated women to work machinery by which the guessing process could be accelerated. By this method the guessed groups rose at once to twenty daily and by the law of increasing returns grew mechanicaly to a maximum of a hundred per day by which time the cypher was approximately readable,

Unfortunately, no documentation available about the “machinery”

The Political Branch solved all German hat codes and their superencipherments

Entropy and Unicity Distance of Genoa 18470 Traffic

- **Key entropy: $H(K) = \log_2 (320! \times (10!)^{320}) = 9{,}150$ bits**
  - Approximately s = 32,000 items, p = 320 pages
  - Assuming we know the division of words to pages but not their order, nor the order of the blocks in the pages – not a realistic assumption!
- **Entropy of the Genoa 18470 "language" – H(L) = 11.3 bits**
  - Measure from a corpus of approx. 23,000 codes
- **Unicity distance:**

  $H(K)/(\log2(s) - H(L)) = 9{,}150/(14.9-11.3) = \underline{2{,}496}$
- **Assuming we can estimate the division of words to pages within a 500 error margin: Unicity distance = $\underline{3{,}278}$**

Entropy and Unicity Distance of a 4-digit Hat Code

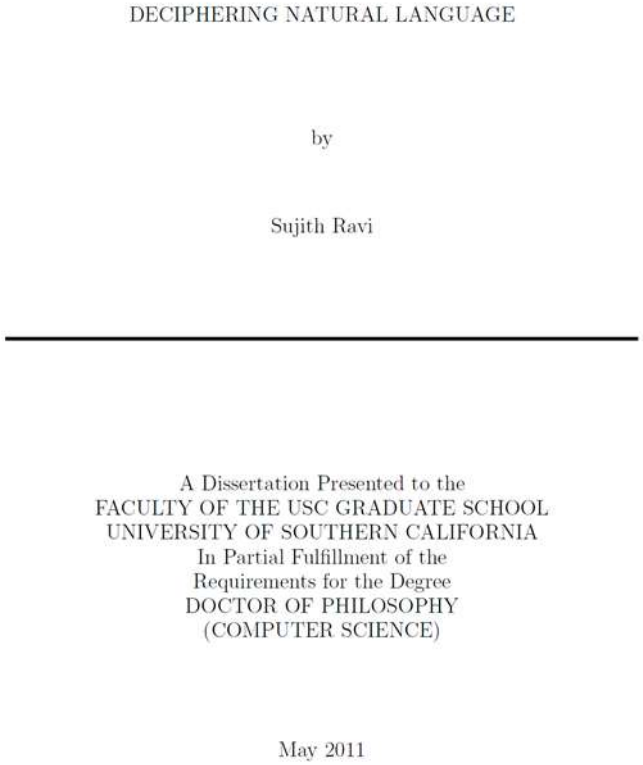
- **Key entropy: $H(K) = \log_2 (10{,}000!) = 118{,}458$**
  - s = 10,000 items
  - Vocabulary fully known (not realistic), order unknown
- **Estimated entropy of the code "language": H(L) = 9.6 bits**
- **Unicity Distance:**

  $H(K)/(\log2(s) - H(L)) = 118{,}458/(13.3\text{-}9.6) = \underline{32{,}122}$

- **Any search algorithm would require very large amounts of material**
  - E.g. 10x the unicity distance = 300K codes
- **The main challenge is the huge key space**
  - In practice, the vocabulary is only partially known → larger H(K)
  - For comparison, H(K) for Enigma cipher is **76**, and H(K) is **logarithmic**

An Interesting Algorithm

- **Sujit Ravi, Ph.D. thesis, 2011**
- **Vocabulary size: 10,000**
  - Assumes vocabulary is known
- **Baysian approach**
  - Optimize $P(w_j/w_i)$ – the probability that word w_j appears after word w_i
  - Large reference corpus of English text
- **Accuracy: 60% to 82%**
- **Compute intensive**



DECIPHERING NATURAL LANGUAGE

by

Sujith Ravi

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

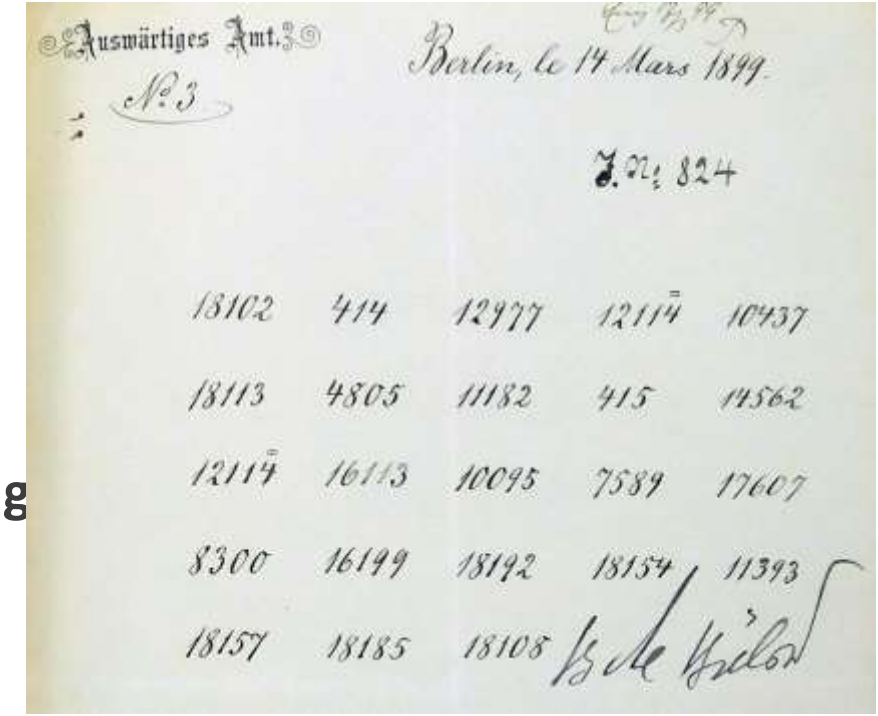
May 2011

An Interesting Algorithm - Illustration

C:	*3894 9411 4357 8446 5433*
O:	a diploma that's good .
D:	a **fence** that's good .
C:	*8593 7932 3627 9166 3671*
O:	three families living here ?
D:	three **brothers** living here ?
C:	*6283 8827 7592 6959 5120 6137 9723 3671*
O:	okay and what did they tell you ?
D:	okay and what did they tell you ?
C:	*9723 3601 5834 5838 3805 4887 7961 9723 3174 4518 9067 4488 9551 7538 7239 9166 3671*
O:	you mean if we come to see you in the afternoon after five you'll be here ?
D:	**i** mean if we come to see you in the afternoon after **thirty** you'll be here ?
	...

Unsolved Diplomatic Codes from WWI – Code 18102

- **Part of the Genoa collection**
- **From 1898 to 1913**
- **Messages with a total of 5,000 codes**
- **Probably an ordered or mixed code**
  - Could be a derivative of 13040
- **No documentation**
  - And no matching plaintext
- **Could use the 18470 corpus as starting point**
- **Same 3-digit codes as for 18470**
  - Numbers and dates


Auswärtiges Amt.

Berlin, le 14 Mars 1899.

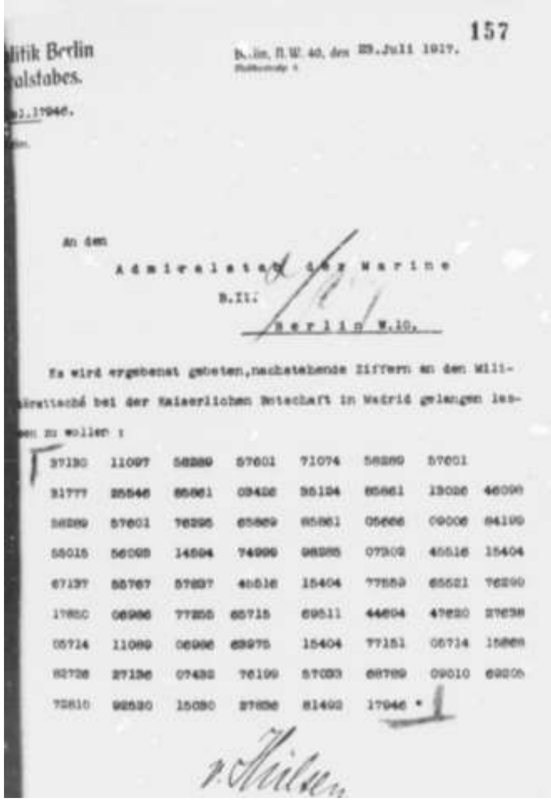
No 3

Z. n. 824

18102	414	12977	12114	10437
18113	4805	11182	415	14562
12114	16113	10095	7589	17607
8300	16199	18192	18154	11393
18157	18185	18108		

Unsolved Naval Attaché Codes from 1917

- **Madrid-Berlin correspondence**
  - Mid-late 1917
  - Hundreds of telegrams
- **Hat codes: 0053, 0064, 0075**
  - No copy available in archives
- **Naval codes**
  - Verkehrsbuch (VB) and Satzbuch (SB)
  - Digital scans available
- **Several superencipherment methods**
  - E.g. sliders, other unknown
- **Solved by Room 40**
  - English transcipts


157

An den

Admiralstab der Marine

B.III

Berlin W.10.

Es wird ergebenst gebeten, nachstehende Ziffern an den Mili-

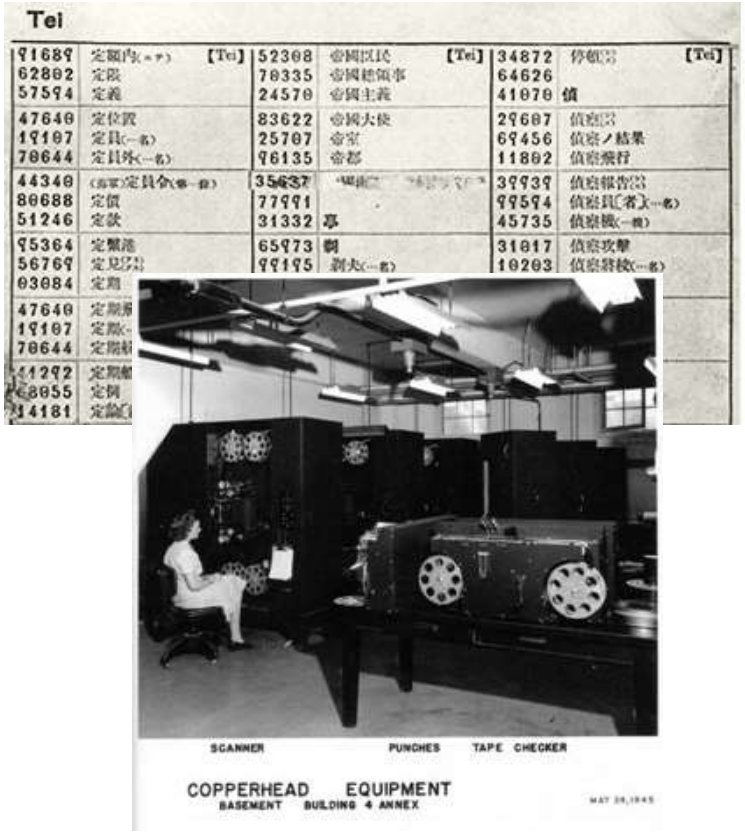
tärattaché bei der Kaiserlichen Botschaft in Madrid gelangen las-

sen zu wollen :

37130 11097 58289 57601 71074 58289 57601

Solving Superencipherments – Computer Algorithms

- **Feasible and effective if method and base codebooks are known**
  - E.g. Verkehrsbuch and Satzbuch
  - Requires transcription of base code book
    - Expensive and long process
- **Also feasible if code values have some mathematical characteristics**
  - E.g. check digit for Japanese JN-25 in WWII
    - Mamba and Copperhead machines
  - Russian Baltic Fleet codes (1930s-1940s)
  - Not applicable for WWI codes



Thank You

June 20, 2018
George Lasry, Ph.D.
University of Kassel, Germany
george.lasry@gmail.com