The Copiale Cipher*

Kevin Knight

USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA, 90292, USA
knight@isi.edu

SDL Language Weaver, Inc.
6060 Center Drive, Suite 150
Los Angeles, CA 90045
kknight@sdl.com

Beáta Megyesi and Christiane Schaefer

Department of Linguistics and Philology
Uppsala University
Box 635, S-751 26 Uppsala, Sweden
beata.megyesi@lingfil.uu.se
christiane.schaefer@lingfil.uu.se

Abstract

The Copiale cipher is a 105-page enciphered book dated 1866. We describe the features of the book and the method by which we deciphered it.

1. Features of the Text

Figure 1 shows a portion of an enciphered book from the East Berlin Academy. The book has the following features:

- It is 105 pages long, containing about 75,000 handwritten characters.
- The handwriting is extremely neat.
- Some characters are Roman letters (such as a and b), while others are abstract symbols (such as ȣ and Δ). Roman letters appear in both uppercase and lowercase forms.
- Lines of text are both left- and right-justified.
- There are only a few author corrections.
- There is no word spacing.

There are no illustrations or chapter breaks, but the text has formatting:

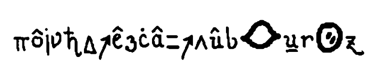
- Paragraphs are indented.
- Some lines are centered.
- Some sections of text contain a double-quote mark (“) before each line.
- Some lines end with full stop (**.**) or colon (**:**). The colon (**:**) is also a frequent word-internal cipher letter.
- Paragraphs and section titles always begin with Roman letters (in capitalized form).

The only non-enciphered inscriptions in the book are “Philipp 1866” and “Copiales 3”, the latter of which we used to name the cipher.

The book also contains preview fragments (“catchwords”) at the bottom of left-hand pages. Each catchword is a copy of the first few letters from the following (right-hand) page. For example, in Figure 1, the short sequence ʒâʌ floats at the bottom of the left page, and the next page begins ʒâʌomı... In early printing, catchwords were used to help printers validate the folding and stacking of pages.

2. Transcription

To get a machine-readable version of the text, we devised the transcription scheme in Figure 2. According to this scheme, the line



is typed as:

```
pi oh j v hd tri arr eh three c. ah
ni arr lam uh b lip uu r o.. zs
```

—

**This material was presented as part of an invited talk at the 4th Workshop on Building and Using Comparable Corpora (BUCC 2011).*

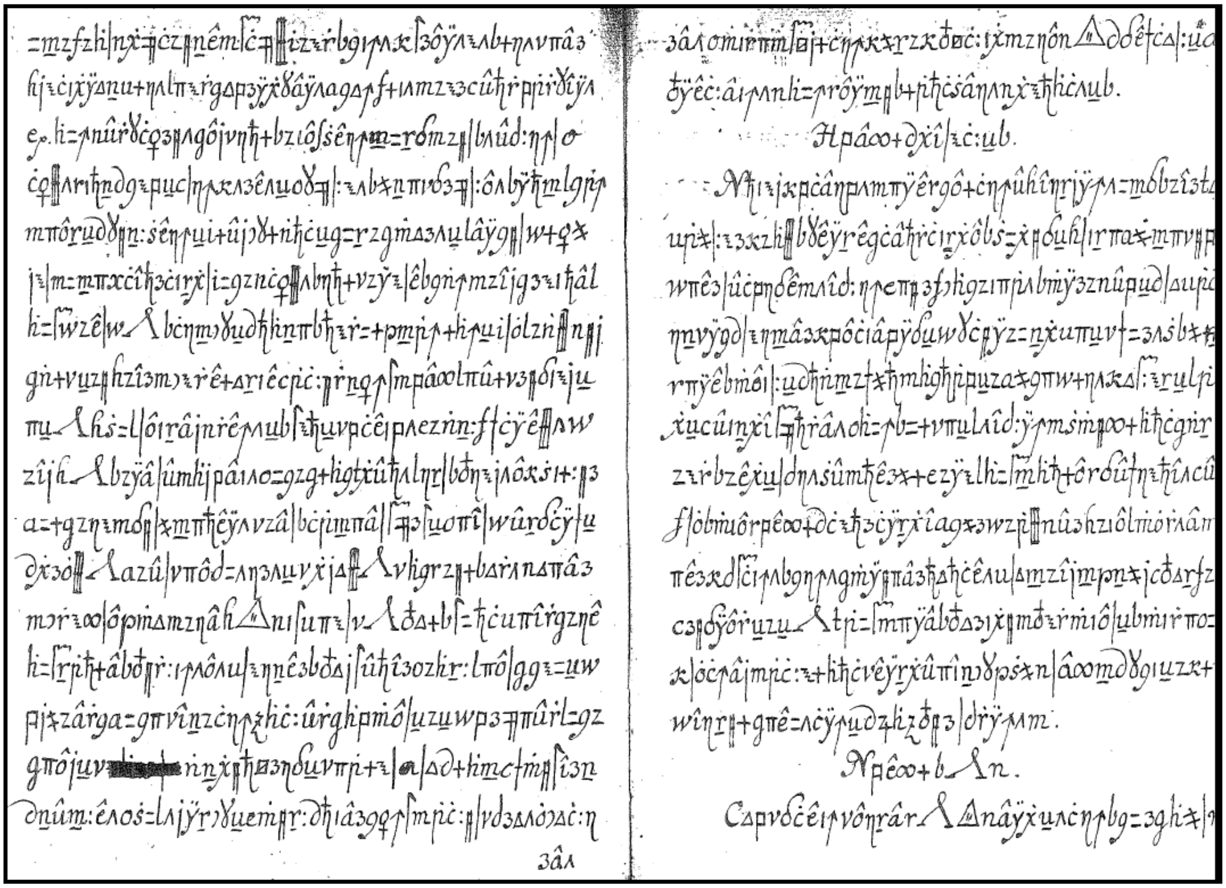


Figure 1. Two pages from the Copiale cipher.

The transcription uses easy-to-reach keyboard characters, so a transcriber can work without taking his/her eyes off the original document.

There are approximately 90 cipher letters, including 26 unaccented Roman letters, a-z. The letters c, h, m, n, p, r, s, and x have dotted forms (e.g., ċ), while the letter i also has an un-dotted form. The letters m, n, r, and u have underlined variants (e.g., m̲), and the vowels have circumflexed variants (e.g., â). The plain letter y does not appear unaccented until page 30, but it appears quite frequently with an umlaut (ÿ).The four Roman letters d, g, n, and z appear in both plain (d, g, n, z) and fancy forms (ə, ℊ, η, ʑ). Capitalized Roman letters are used to start paragraphs. We transcribe these with A-Z, though we down-case them before counting frequencies (Section 3). Down-casing D, G, N, and Z is not trivial, due to the presence of both plain and fancy lowercase forms.

The non-Roman characters are an eclectic mix of symbols, including some Greek letters. Eight symbols are rendered larger than others in the text: ⅄, ⊙, Δ, ✕, ◇, ∮, ⌓, and Π.

We transcribed a total of 16 pages (10,840 letters). We carried out our analysis on those pages, after stripping catchwords and down-casing all Roman letters.

3. Letter Frequencies and Contexts

Figure 3 shows cipher letter frequencies. The distribution is far from flat, the most frequent letter being ^ (occurring 412 times). Here are the most common cipher digraphs (letter pairs) and trigraphs, with frequencies:

ɔ ħ 99 — ɔ ħ ʌ 47
ċ : 66 — ċ : u̲ 23
ħ ʌ 49 — η ɔ ħ 22
: u̲ 48 — ÿ ɔ ħ 18
z ǁ 44 — ɦ ċ | 17

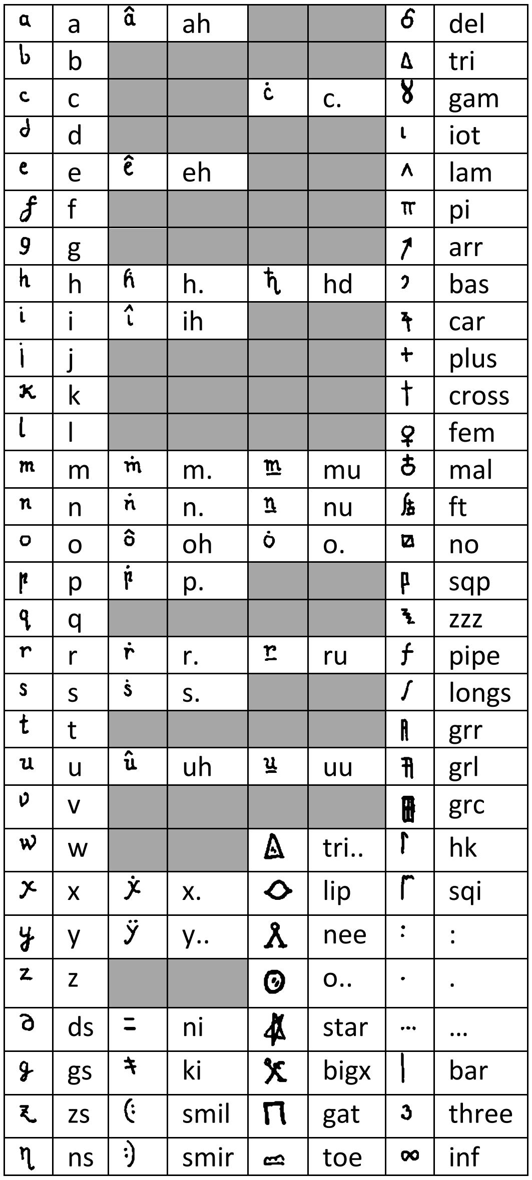


Figure 2. Transcription scheme. Columns alternate between the cipher letters and their transcriptions.

The full digraph counts reveal interesting patterns among groups of letters. For example, letters with circumflexes (â, ê, î, ô, û) have behaviors in common: all five tend to be preceded by z and π, and all five tend to be followed by ʒ and ¡. To get a better handle on letter similarities, we automatically clustered the cipher letters based on their contexts. The result is shown in Figure 4. We did the clustering as follows. For each distinct letter x, we created a

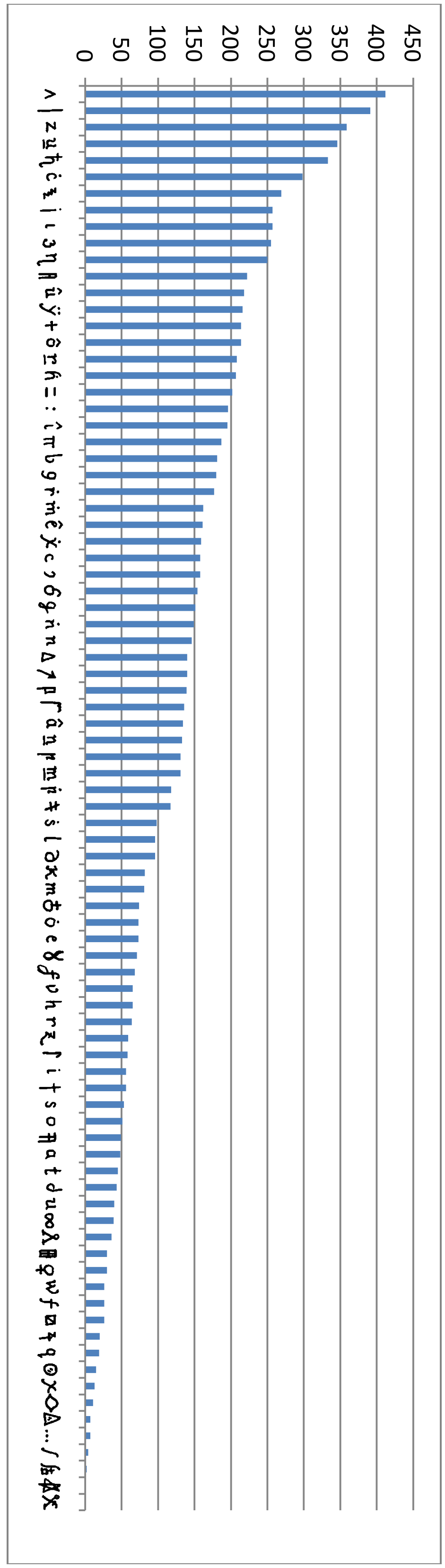


Figure 3. Cipher letter frequencies.

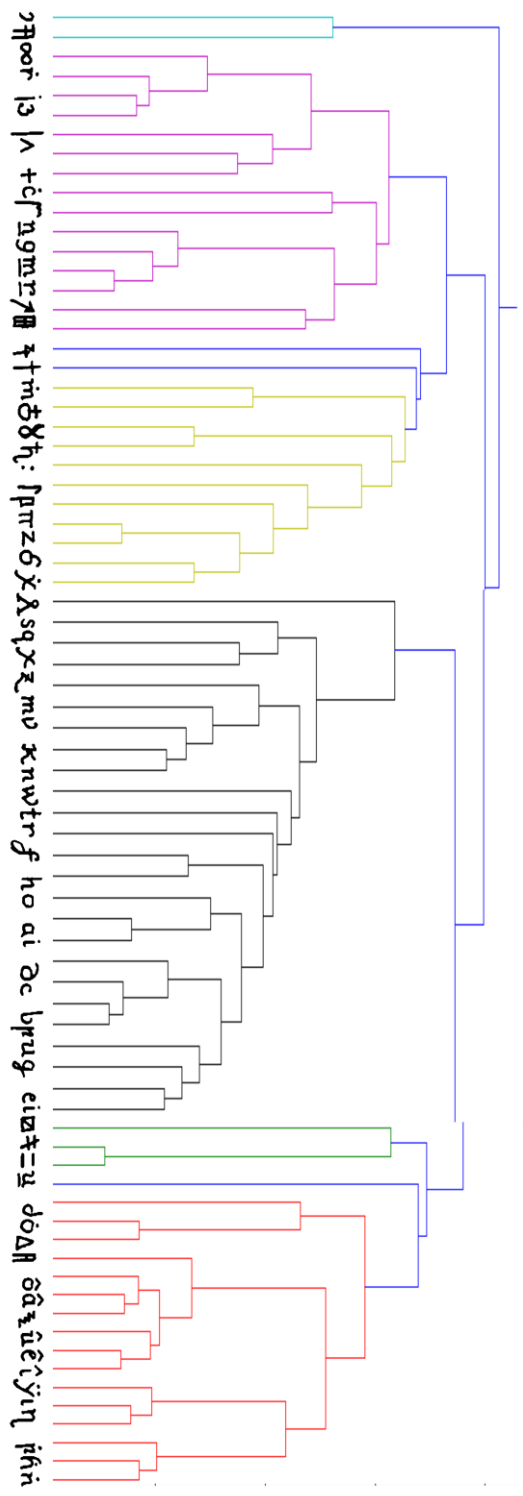


Figure 4. Automatic clustering of cipher letters based on similarity of contexts.

co-occurrence vector of length 90, to capture the distribution of letters than precede x. For example, if x is preceded 12 times by π, 0 times by û, 4 times by ÿ, 1 time by δ, etc, then its vector looks like this: [12, 0, 4, 1, …]. For the same letter x, we created another vector that captures the distribution of letters than follow x, e.g., [0, 0, 7, 2, …]. Then we concatenated the two vectors to create v(x) = [12, 0, 4, 1, …, 0, 0, 7, 2, …]. We deemed two letters a and b to be similar if the cosine distance between v(a) and v(b) is small, indicating that they appear in similar contexts. We used the Scipy software (http://users.soe.ucsc.edu/~eads/cluster.html) to perform and plot a clustering that incrementally merges similar letters (and groups of letters) in a bottom-up fashion.

The cluster diagram confirms that circumflexed letters (â, ê, î, ô, û) behave similarly. It also shows that the unaccented Roman letters form a natural grouping, as do underlined letters. Merges that happen low in the cluster map indicate very high similarity, e.g., the group (ÿ, ι, η).

4. First Decipherment Approach

Building on the self-similarity of Roman letters, our first theory was that the Roman letters carry all the information in the cipher, and that all other symbols are NULLs (meaningless tokens added after encipherment to confuse cryptanalysis). If we remove all other symbols, the remaining Roman letters indeed follow a typical natural language distribution, with the most popular letter occurring 12% of the time, and the least popular letters occurring rarely.

The revealed sequence of Roman letters is itself nonsensical, so we posited a simple substitution cipher. We carried out automatic computer attacks against the revealed Roman-letter sequence, first assuming German source, then English, then Latin, then forty other candidate European and non-European languages. The attack method is given in [Knight et al, 2006]. That method automatically combines plaintext-language identification with decipherment. Unfortunately, this failed, as no

language identified itself as a more likely plaintext candidate than the others.

We then gave up our theory regarding NULLs and posited a homophonic cipher, with each plaintext letter being encipherable by any of several distinct cipher letters. While a well-executed homophonic cipher will employ a flat letter frequency distribution, to confound analysis, we guessed that the Copiale cipher is not optimized in this regard.

We confirmed that our computer attack does in fact work on a synthetic homophonic cipher, i.e., it correctly identifies the plaintext language, and yields a reasonable, if imperfect, decipherment. We then loosed the same attack on the Copiale cipher. Unfortunately, all resulting decipherments were nonsense, though there was a very slight numerical preference for German as a candidate plaintext language.

5. Second Decipherment Approach

We next decided to focus on German as the most likely plaintext language, for three reasons:

- the book is located in Germany
- the computer homophonic attack gave a very slight preference to German
- the book ends with the inscription "Philipp 1866", using the German double-p spelling.

Pursuing the homophonic theory, our thought was that all five circumflexed letters (â, ê, î, ô, û), behaving similarly, might represent the same German letter. But which German letter? Since the circumflexed letters are preceded by z and π, the circumflexed letters would correspond to the German letter that often follows *whatever z and π stand for*. But what do they, in turn, stand for?

From German text, we built a digraph frequency table, whose the most striking characteristic is that C is almost always followed by H. The German CH pair is similar to the English QU pair, but C is fairly frequent in German. A similar digraph table for the cipher letters shows that ɔ is almost always followed by ɬ. So we posited our first two substitutions: ɔ=C and ɬ=H. We then looked for what typically precedes and follows CH in German, and what typically precedes and follows ɔɬ in the cipher. For example, ɔɬʌ is the most frequent cipher trigraph, while CHT is a common German trigraph. We thus hypothesized the further substitution ʌ=T, and this led to a cascade of others. We retracted any hypothesis that resulted in poor German digraphs and trigraphs, and in this way, we could make steady progress (Figure 5).

The cluster map in Figure 4 was of great help. For example, once we established a substitution like ÿ=I, we could immediately add ɳ=I and ɪ=I, because the three cipher letters behave so similarly. In this way, we mapped all circumflexed letters (â, ê, î, ô, û) to plaintext E. These leaps were frequently correct, and we soon had substitutions for over 50 cipher letters.

Despite progress, some very frequent German trigraphs like SCH were still drastically under-represented in our decipherment. Also, many cipher letters (including all unaccented Roman letters) still lacked substitution values. A fragment of the decipherment thus far looked like this (where "?" stands for an as-yet-unmapped cipher letter):

?GEHEIMER?UNTERLIST?VOR?DIE?GESELLE
?ERDER?TITUL
?CEREMONIE?DER?AUFNAHME

On the last line, we recognized the two words CEREMONIE and DER separated by a cipher letter. It became clear that *the unaccented Roman letters serve as spaces in the cipher*. Note that this is the opposite of our first decipherment approach (Section 4). The non-Roman letters are not NULLs -- they carry virtually all the information. This also explains why paragraphs start with capitalized Roman letters, which look nice, but are meaningless.

We next put our hypothesized decipherment into an automatic German-to-English translator (www.freetranslation.com), where we observed that many plaintext words were still untranslatable. For example, ABSCHNITL was not recognized as a translatable German word. The final cipher letter for this word is colon (:), which we had mapped previously to L. By replacing the final L in ABSCHNITL with various letters of the alphabet (A-Z), we hit on the recognized word

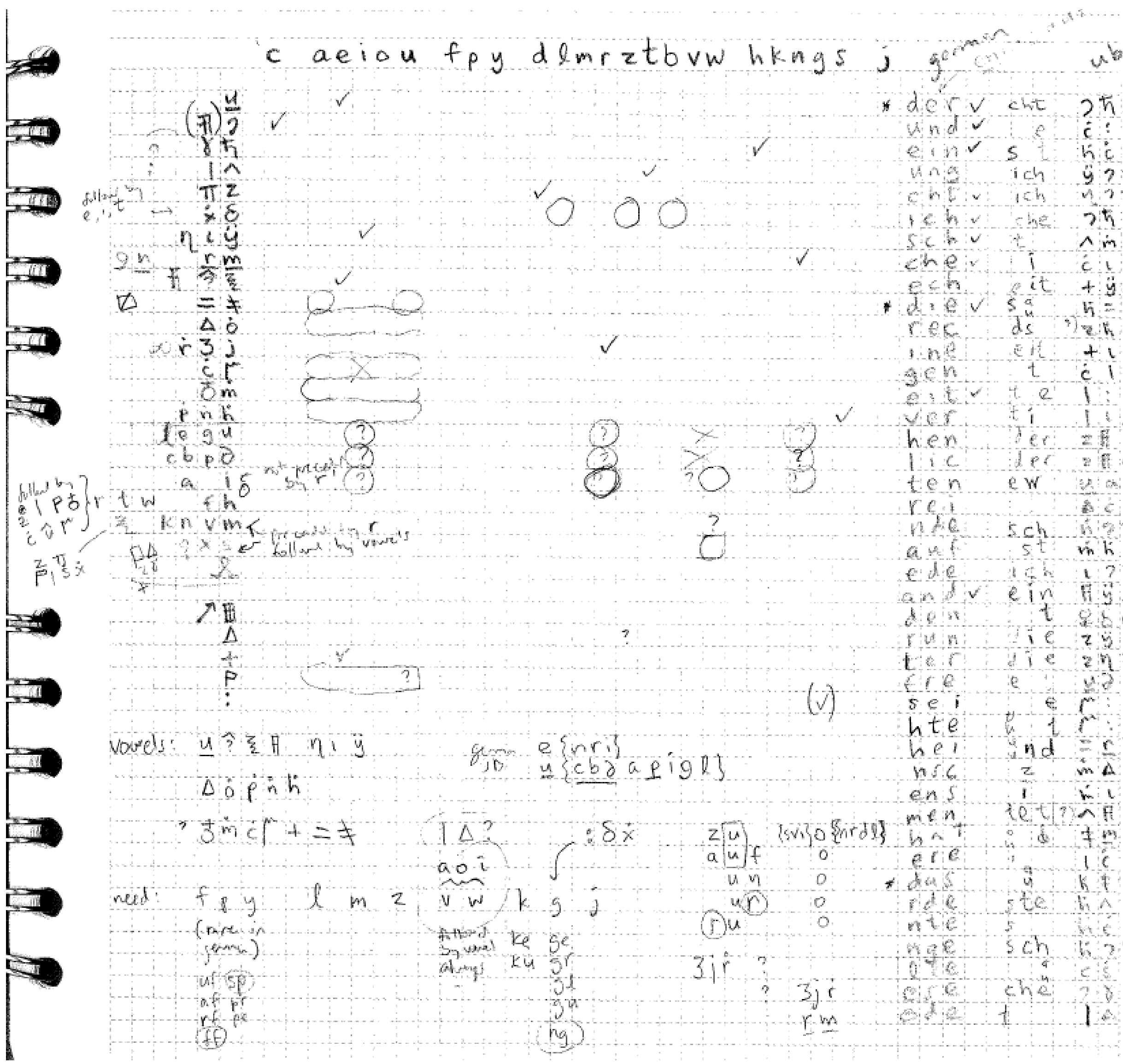
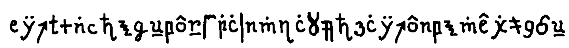


Figure 5. Progress of decipherment. The main grid shows plaintext (German) letters across the top and ciphertext letters down the left side. The ciphertext letters are grouped into clusters. To the right of the main grid are frequent German trigraphs (der, und, ein, …) and frequent cipher trigraphs (ↄħʌ, ċ:u̲, ĥċ|, …), with the two columns being separated by hypothesized trigraph decipherments.

ABSCHNITT (translated as “section”). We then realized that the function of colon (:) is to double the previous consonant (whether it be T, L, F, or some other letter). Old German writing uses a stroke with the same function.

The cipher letter † was still unknown, appearing in partially deciphered words like †AFLNER, †NUPFTUCHS, and GESELL†AFLT. We tried substituting each of the letters A-Z for †, but this did not yield valid German. However, we found GESELLSCHAFLT in a German dictionary, so we concluded that † stands for SCH. This opened the door for other multi-plaintext-letter substitutions.

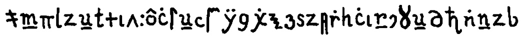
Finally, we brought full native-German expertise to bear, by taking hypothesized decipherments (hyp) and correcting them (corr):

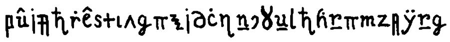


hyp: is mache ebenfals wilhuhrlise bewegunge
corr: i**ch** mache ebenfals wil**kü**hrli**ch**e bewegunge**n**


hyp: dos mit der andern hand
corr: do**ch** mit der andern hand


hyp: dritlens einer n mlt tobach mit de daume
corr: drit**t**ens einer **? ???** tobach mit de**m** daum**en**


hyp: und de mitlelde finger der linche hand
corr: und de**m** mit**t**el**st**en finger der linche**n** hand


hyp: beruhre mit der linche hand dein
corr: ber**ü**hre mit der linche**n** hand dein

This allowed us to virtually complete our table of substitutions (Figure 6). Three cipher letters remained ambiguous:

- ▓ could represent either SS or S
- ɣ could represent either H or K
- ʉ could represent either EN or EM

However, these three symbols are ambiguous only with respect to deciphering into modern German, not into old German, which used different spelling conventions.

The only remaining undeciphered symbols were the large ones: λ, ⊙, Δ, ✗, ◇, ∦, ≃, and Π. These appear to be logograms, standing for the names of (doubly secret) people and organizations, as we see in this section: "the λ asks him whether he desires to be ◇".

Plaintext (German)	Ciphertext
A	ṗ ṅ ĥ ♀*
Ä	♀*
B	ꝑ
C	ɔ
D	π z
E	â ê î ô û ꞵ ʑ
F	ɼ
G	6 ẋ
H	ђ ɣ*
I	ÿ η ι
J	ɟ
K	ɣ*
L	ċ
M	+
N	m̲ r̲ n̲ g
O	Δ ȯ
Ö	⊠
P	∂
R	ṙ 3 ¦
S	ǀ ▓*
T	ʌ
U	= ɟ
Ü	ꞵ̈
V	ȣ
W	ṁ
X	ʄ
Y	∞
Z	ṡ
SCH	†
SS	▓*
ST	ɼ
CH	↗
repeat previous consonant	:
EN / EM	ʉ
space	a b c ∂ e f ꞵ g h i x l m n o p q r s ſ t u v w x y z

Figure 6. Letter substitutions resulting from decipherment. Asterisks (*) indicate ambiguous cipher letters that appear twice in the chart. This table does not include the large characters: λ, ⊙, Δ, ✗, ◇, ∦, ≃, and Π.

6. Contents

The book describes the initiation of "DER CANDIDAT" into a secret society, some functions of which are encoded with logograms. Appendix A contains our decipherment of the beginning of the manuscript.

7. Conclusion

We described the Copiale cipher and its decipherment. It remains to transcribe the rest of the manuscript and to do a careful translation. The document may contain further encoded information, and given the amount of work it represents, we believe there may be other documents using the same or similar schemes.

8. Acknowledgments

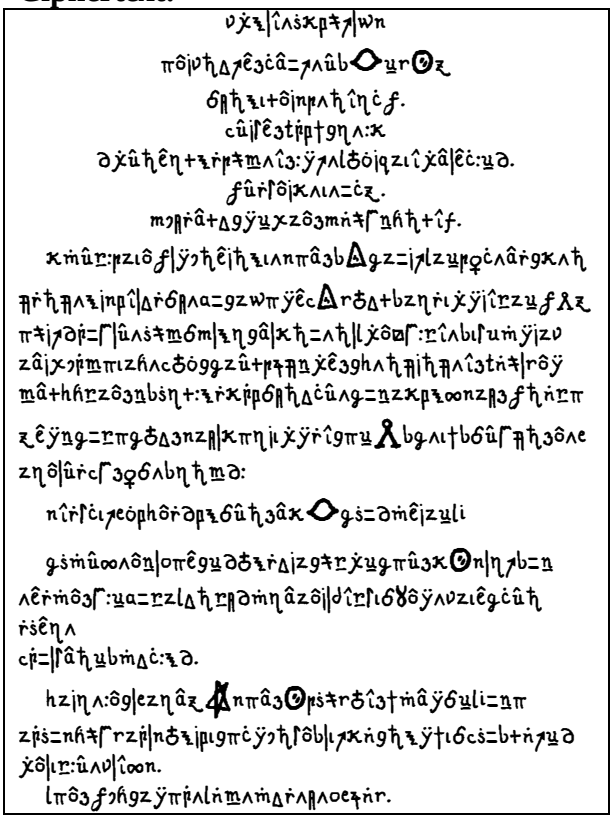
Thanks to Jonathan Graehl for plotting the cluster map of cipher letters. This work was supported in part by NSF grant 0904684.

9. References

K. Knight, A. Nair, N. Rathod, and K. Yamada, "Unsupervised Analysis for Decipherment Problems", Proc. ACL-COLING, 2006.

Appendix A

Ciphertext:



Plaintext, as deciphered:

gesetz buchs
der hocherleuchte ⏀ e ⓪
geheimer theil.
erster abschnitt
geheimer unterricht vor die gesellen.
erster titul.
ceremonien der aufnahme.

wenn die sicherheit der 🅰 durch den ältern thürheter besorget und die 🅰 vom dirigirenden 🧍 durch aufsetzung seines huths geöffnet ist wird der candidat von dem jüngern thürhüter aus einem andern zimmer abgeholet und bey der hand ein und vor des dirigirenden 🧍 tisch geführet dieser frägt ihn:

erstlich ob er begehre ⏀ zu werden

zweytens denen verordnungen der ⓪ sich unterwerffen und ohne wiederspenstigkeit die lehrzeit ausstehen wolle.

drittens die 🕯 der ⓪ gu verschweigen und dazu auf das verbindlichste sich anheischig zu machen gesinnet sey.

der candidat antwortet ja.

Initial translation:

First lawbook
of the ⏀ e ⓪
Secret part.
First section
Secret teachings for apprentices.
First title.
Initiation rite.

If the safety of the 🅰 is guaranteed, and the 🅰 is opened by the chief 🧍, by putting on his hat, the candidate is fetched from another room by the younger doorman and by the hand is led in and to the table of the chief 🧍, who asks him:

First, if he desires to become ⏀.

Secondly, if he submits to the rules of the ⓪ and without rebelliousness suffer through the time of apprenticeship.

Thirdly, be silent about the 🕯 of the ⓪ and furthermore be willing to offer himself to volunteer in the most committed way.

The candidate answers yes.