

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
knight@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 **ϣ** 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.

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

- 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 **ϣâλ** floats at the bottom of the left page, and the next page begins **ϣâλomi...** 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

πδ|ρϣΔ|ϣεζcâ=|λûb|ϣr|ϣε

is typed as:

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

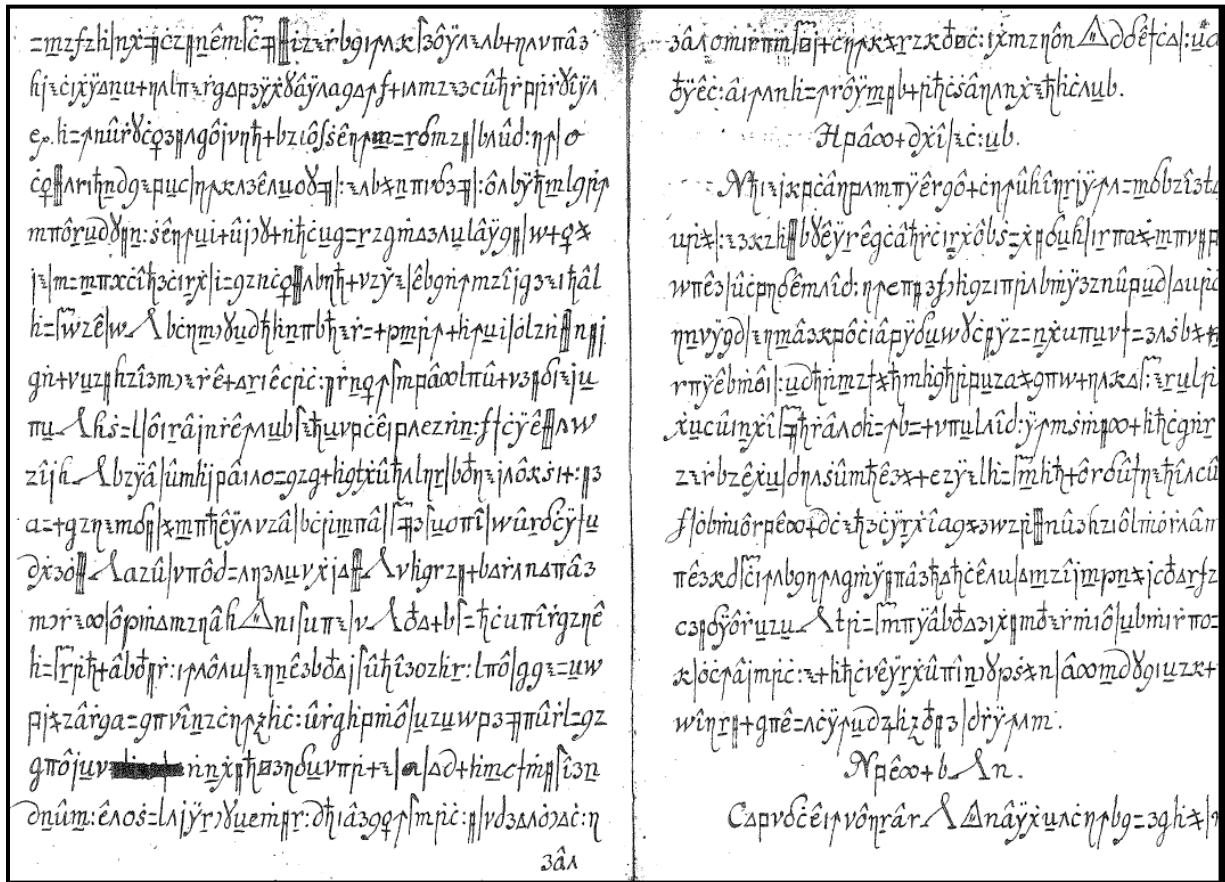


Figure 1. Two pages from the Copiale cipher.

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., \dot{c}), while the letter i also has an un-dotted form. The letters m, n, r, and u have underlined variants (e.g., \underline{m}), and the vowels have circumflexed variants (e.g., \hat{a}). The plain letter y does not appear unaccented until page 30, but it appears quite frequently with an umlaut (\ddot{y}). The four Roman letters d, g, n, and z appear in both plain (\mathfrak{d} , \mathfrak{g} , \mathfrak{n} , \mathfrak{z}) and fancy forms (\mathfrak{d} , \mathfrak{g} , \mathfrak{n} , \mathfrak{z}). 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: Λ , Θ , Δ , \times , \circ , \mathfrak{A} , ∞ , 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 \wedge (occurring 412 times). Here are the most common cipher digraphs (letter pairs) and trigraphs, with frequencies:

\mathfrak{c} \mathfrak{h}	99	\mathfrak{c} \mathfrak{h} \wedge	47
\dot{c} :	66	\dot{c} : \underline{u}	23
\mathfrak{h} \wedge	49	η \mathfrak{c} \mathfrak{h}	22
: \underline{u}	48	\ddot{y} \mathfrak{c} \mathfrak{h}	18
\mathfrak{z} \mathfrak{h}	44	\mathfrak{h} \dot{c}	17

a	a	â	ah			δ	del
b	b					Δ	tri
c	c			ç	c.	ϝ	gam
d	d					ι	iot
e	e	ê	eh			^	lam
f	f					π	pi
g	g					ʄ	arr
h	h	ĥ	h.	h̃	hd	ʘ	bas
i	i	î	ih			ʔ	car
j	j					+	plus
k	k					†	cross
l	l					♀	fem
m	m	m̂	m.	m̃	mu	♁	mal
n	n	n̂	n.	ñ	nu	♁	ft
o	o	ô	oh	ó	o.	⊞	no
p	p	p̂	p.			♁	sqp
q	q					z	zzz
r	r	r̂	r.	r̃	ru	f	pipe
s	s	ŝ	s.			f	longs
t	t					fl	grr
u	u	û	uh	ũ	uu	fl	grl
v	v					fl	grc
w	w				Δ	tri..	hk
x	x	x̂	x.	⊙	lip	fl	sqi
y	y	ŷ	y..	λ	nee	:	:
z	z			⊙	o..	.	.
ð	ds	=	ni	♁	star
g	gs	κ	ki	κ	bigx		bar
z	zs	ε	smil	Π	gat	ʘ	three
η	ns	;)̂	smir	∞	toe	∞	inf

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 ʘ and j. 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

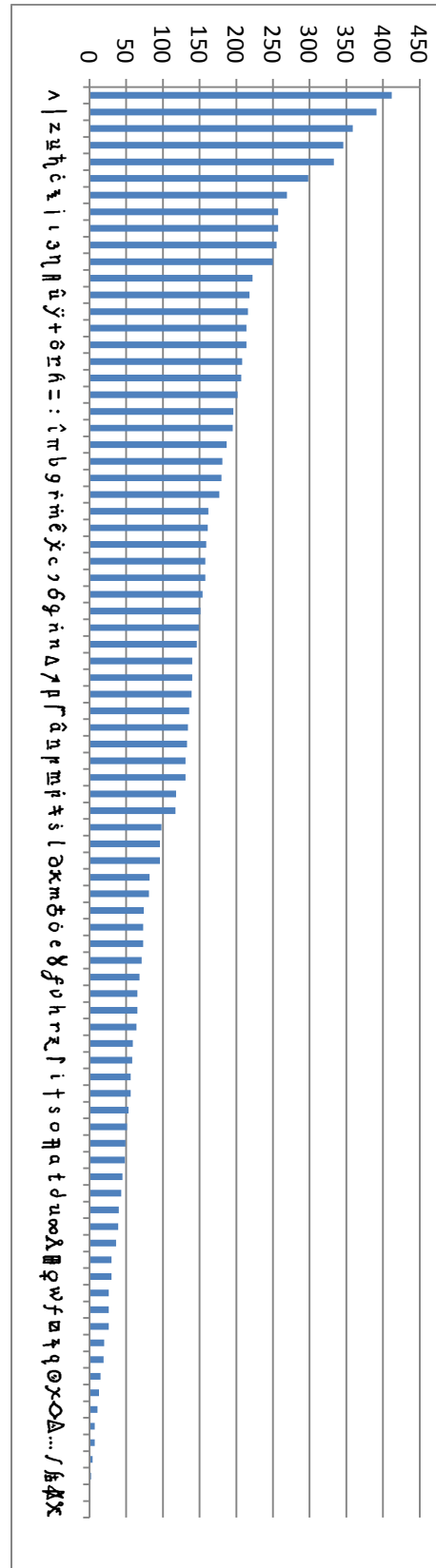


Figure 3. Cipher letter frequencies.

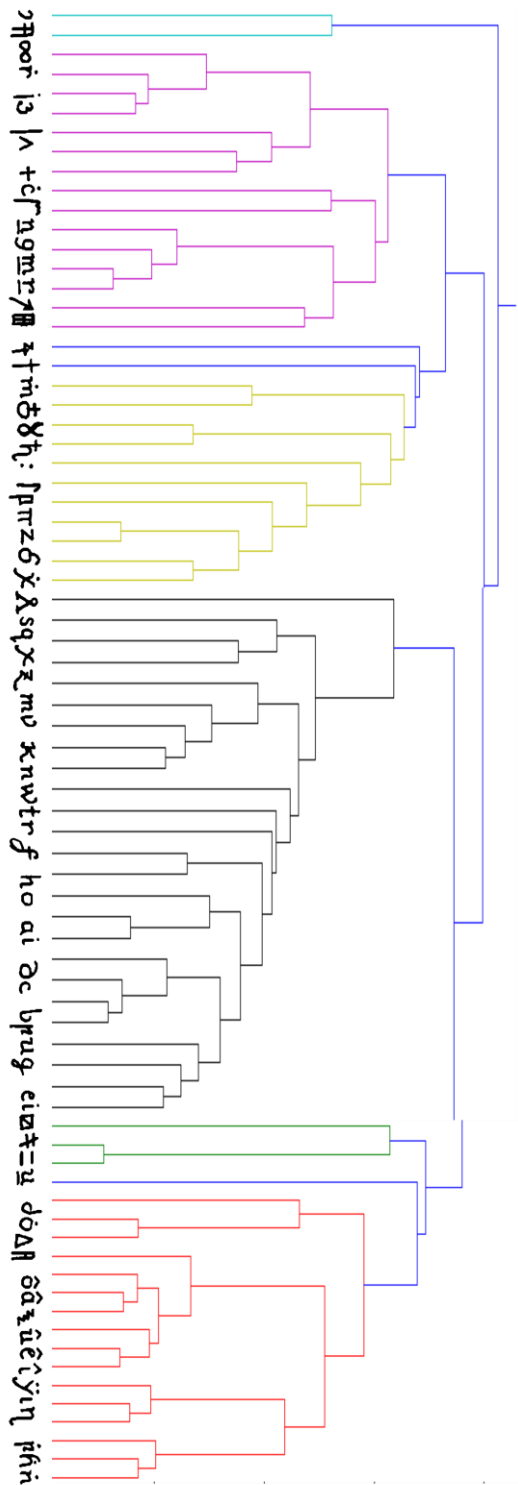


Figure 4. Automatic clustering of cipher letters based on similarity of contexts.

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 \hat{u} , 4 times by \check{y} , 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 $v(x) = [12, 0, 4, 1, \dots, 0, 0, 7, 2, \dots]$. 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 (\hat{a} , \hat{e} , \hat{i} , \hat{o} , \hat{u}) 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 (\check{y} , 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 (\hat{a} , \hat{e} , \hat{i} , \hat{o} , \hat{u}), 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 pi 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 γ is almost always followed by \mathfrak{h} . So we posited our first two substitutions: $\gamma=C$ and $\mathfrak{h}=H$. We then looked for what typically precedes and follows CH in German, and what typically precedes and follows $\gamma\mathfrak{h}$ in the cipher.

For example, $\gamma\mathfrak{h}\lambda$ is the most frequent cipher trigraph, while CHT is a common German trigraph. We thus hypothesized the further substitution $\lambda=T$, 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 $\mathfrak{y}=I$, we could immediately add $\eta=I$ and $\iota=I$, because the three cipher letters behave so similarly. In this way, we mapped all circumflexed letters (\hat{a} , \hat{e} , \hat{i} , \hat{o} , \hat{u}) 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 trigrams (der, und, ein, ...) and frequent cipher trigrams (ɔhΛ, c:u, hcl, ...), with the two columns being separated by hypothesized trigram 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):

eɔʔt+nchzɔpɔrɯic|nriηcδʔhɔcɔʔɔnpzmeɔʔgɔu
 hyp: is mache ebenfals wiluhrlise bewegunge
 corr: ich mache ebenfals wilkührliche bewegungen

απΔβ+ιλγζηθρσϕχψζεζηθπσ
 hyp: dos mit der andern hand
 corr: doch mit der andern hand

ρζιηλ:ίγ|εϋñî|εμη+:αβλὸρπῆδρ+ιλφπυγζή=+υε
 hyp: dritlens einer n mlt tobach mit de daume
 corr: drittens einer ??? tobach mit dem daumen

κμπ(ζυτ+ιλ:δέ|υε|γῶχξζσζρῆθίρσδϋθῆαυζβ
 hyp: und de mittelde finger der linche hand
 corr: und dem mittelsten finger der linchen hand

ρû|ηῆ|ρῆς+ιλγπξ|δὲηυδϋλῆ|ρπμζρῶγρ
 hyp: beruhre mit der linche hand dein
 corr: berühre mit der linchen 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
- ⚡ could represent either H or K
- ϋ 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: ⚡, ⚡, ⚡, ⚡, ⚡, ⚡, and ⚡. These appear to be logograms, standing for the names of (doubly secret) people and organizations, as we see in this section: “the ⚡ asks him whether he desires to be ⚡”.

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.

Plaintext (German)	Ciphertext
A	î ñ ã ϕ*
Ä	ϕ*
B	β
C	γ
D	π ζ
E	â ê î ô û ρ ξ
F	ϒ
G	δ χ
H	η ⚡*
I	ÿ η ι
J	τ
K	⚡*
L	é
M	+
N	μ ρ υ ϑ
O	Δ ó
Ö	⚡
P	δ
R	ř 3 i
S	⚡*
T	^
U	= ð
Ü	ϕ
V	θ
W	ñ
X	ƒ
Y	∞
Z	š
SCH	†
SS	⚡*
ST	†
CH	↑
repeat previous consonant	:
EN / EM	ϋ
space	a b c d e f g h i k l m n o p q r 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:

⚡, ⚡, ⚡, ⚡, ⚡, ⚡, and ⚡.

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:

lit:mzq|bl
 vχx|l̄s̄k̄p̄k̄jw̄n
 π̄ōp̄h̄Δ̄r̄ē̄z̄c̄â̄=̄r̄l̄û̄b̄⊕̄ūr̄⊕̄z̄
 Ⓛ̄h̄h̄z̄i+Ⓛ̄īn̄r̄l̄h̄î̄ñ̄c̄f̄.
 c̄û̄j̄f̄ē̄z̄t̄p̄t̄ḡη̄λ̄:κ̄
 Ⓛ̄x̄û̄h̄ē̄η̄+κ̄r̄k̄m̄l̄î̄z̄:γ̄r̄l̄ō̄j̄q̄z̄īx̄â̄|ē̄c̄:ūd̄.
 f̄û̄r̄f̄ō̄j̄k̄l̄l̄l̄c̄z̄.
 m̄j̄r̄â̄+Δ̄ḡȳx̄z̄ō̄z̄m̄î̄k̄f̄ḡh̄h̄+î̄f̄.
 κ̄m̄û̄r̄:ρ̄z̄īō̄f̄j̄ȳj̄h̄ē̄j̄h̄z̄īl̄n̄π̄â̄z̄b̄Δ̄ḡz̄=̄j̄r̄z̄ūr̄q̄c̄l̄â̄r̄ḡκ̄λ̄h̄
 η̄r̄h̄η̄λ̄z̄j̄n̄r̄l̄Δ̄r̄ō̄r̄l̄ā=̄ḡz̄w̄π̄ȳē̄c̄Δ̄r̄ō̄Δ̄+̄b̄z̄η̄r̄īx̄ȳj̄î̄r̄z̄ūf̄λ̄z̄
 π̄k̄j̄r̄ō̄r̄=̄f̄|û̄l̄s̄k̄m̄ō̄m̄|z̄η̄ḡâ̄|κ̄h̄=̄λ̄h̄|l̄x̄ō̄w̄f̄:π̄l̄b̄īūm̄ȳj̄z̄v̄
 z̄â̄j̄x̄j̄m̄π̄z̄h̄l̄c̄ō̄ō̄ḡz̄û̄+̄r̄k̄η̄z̄ē̄z̄ḡh̄l̄h̄η̄h̄η̄l̄z̄t̄î̄k̄|r̄ō̄ȳ
 m̄â̄+̄h̄h̄r̄z̄ō̄z̄û̄b̄s̄η̄+̄z̄r̄k̄r̄ō̄r̄h̄h̄Δ̄c̄û̄λ̄ḡ=̄ūz̄κ̄r̄z̄ōōn̄z̄r̄z̄f̄h̄î̄r̄π̄
 z̄ē̄ȳūḡ=̄r̄π̄ḡō̄Δ̄z̄n̄z̄r̄l̄κ̄π̄η̄j̄l̄x̄ȳr̄î̄ḡπ̄ūλ̄b̄ḡl̄īt̄b̄ō̄û̄f̄η̄h̄z̄ō̄l̄ē
 z̄η̄ō̄j̄r̄c̄f̄z̄q̄ō̄l̄b̄η̄h̄m̄d̄:
 n̄î̄r̄f̄c̄īr̄ēō̄p̄h̄ō̄r̄ō̄r̄z̄ō̄û̄h̄z̄â̄κ̄⊕̄ḡs̄=̄ō̄m̄ē̄j̄z̄ūl̄ī
 ḡs̄m̄û̄ōōl̄ō̄n̄|ōπ̄ē̄ḡȳd̄ō̄z̄r̄Δ̄j̄z̄ḡk̄r̄x̄ūḡπ̄û̄z̄κ̄⊕̄η̄j̄r̄b̄=̄ū
 λ̄ē̄r̄m̄ō̄z̄f̄:ūā=̄r̄z̄l̄Δ̄h̄r̄r̄ō̄m̄η̄â̄z̄ō̄j̄|ō̄î̄r̄l̄īō̄l̄ō̄ȳl̄z̄īē̄ḡc̄û̄h̄
 r̄s̄ē̄η̄λ̄
 c̄f̄=̄f̄f̄â̄h̄z̄b̄m̄Δ̄c̄:z̄d̄.
 h̄z̄j̄η̄λ̄:ō̄ḡ|ēz̄η̄â̄z̄λ̄n̄π̄â̄z̄⊕̄r̄s̄k̄r̄ō̄î̄z̄t̄m̄â̄ȳō̄z̄l̄ī=̄ūπ̄
 z̄r̄s̄=̄n̄h̄k̄f̄r̄z̄r̄|n̄ō̄z̄j̄r̄īḡπ̄c̄ȳj̄h̄f̄ō̄b̄|r̄k̄h̄ḡh̄z̄ȳt̄īō̄c̄s̄=̄b̄t̄h̄r̄z̄d̄
 x̄ō̄j̄r̄:û̄l̄v̄|î̄ōōn̄.
 l̄π̄ō̄z̄f̄j̄h̄ḡz̄ȳπ̄r̄l̄î̄m̄l̄m̄Δ̄r̄l̄r̄l̄ōēz̄h̄r̄.

Plaintext, as deciphered:

gesetz buchs
 der hocherleuchte ⊕ e ⊕
 geheimer theil.
 erster abschnitt
 geheimer unterricht vor die gesellen.
 erster titul.
 ceremonien der aufnahme.
 wenn die sicherheit der Δ durch den ältern
 thürhuter besorget und die Δ vom dirigirenden λ
 durch aufsetzung seines huths geöffnet ist wird der
 candidat von dem jüngern thürhüter aus einem andern
 zimmer abgeholt und bey der hand ein und vor des
 dirigirenden λ tisch geführt dieser fragt ihn:
 erstlich ob er begehre ⊕ zu werden
 zweytens denen verordnungen der ⊕ sich
 unterwerffen und ohne widerspenstigkeit die lehrzeit
 ausstehen wolle.
 drittens die λ der ⊕ gu verschweigen und dazu
 auf das verbindlichste sich anheischig zu machen
 gesinnet sey.
 der candidat antwortet ja.

Initial translation:

First lawbook
 of the ⊕ e ⊕
 Secret part.
 First section
 Secret teachings for apprentices.
 First title.
 Initiation rite.
 If the safety of the Δ is guaranteed, and the Δ is
 opened by the chief λ, 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 λ, who asks him:
 First, if he desires to become ⊕.
 Secondly, if he submits to the rules of the ⊕ and
 without rebelliousness suffer through the time of
 apprenticeship.
 Thirdly, be silent about the λ of the ⊕ and
 furthermore be willing to offer himself to volunteer
 in the most committed way.
 The candidate answers yes.