.92

Reading With-out a Key

[illegible]

[illegible]

[illegible]

[illegible]

Reading Without a Key

THE VOYNICH MANUSCRIPT, EXAMINED

What the object, the script and the numbers can tell us about a book no one has read in six hundred years

Kyle Prier

with

iv

READING

Claude (Anthropic)

Copyright © 2026 Kyle Prier. Working draft, version 1.0, October 2026.

All images of Beinecke MS 408 are reproduced from the digital

facsimile published by the Beinecke Rare Book & Manuscript Library, Yale University (`https://collections.library.yale.edu/catalog/2002046`). The manuscript is in the public domain.

All statistics in this book were computed from the Zandbergen–Landini transliteration (ZL3b, 13 May 2025, IVTFF 2.0) distributed at `https://www.voynich.nu`, with the scripts printed in Appendix A. Figures were produced with those scripts.

This book was written by a human author working with an AI system. Appendix D describes how, and what each contributed.

Set in Spectral and Inter. Typeset with XeLaTeX.

Contents

Prologue: the first page vii

I The Object 1

1 A Book Nobody Can Read 2

2 Skin, Ink and Pigment 6

3 Where It Has Been 10

4 In What Order? 14

II The Script 18

5 Writing It Down 19

6 Every Letter Knows Its Place 23

7 The Line and the Paragraph 27

8 Two Languages, Five Hands 31

9 The Numbers 36

III The Pictures 41

10 Plants 42

11 Stars and the Zodiac 47

12 The Bathers 53

13 The Rosettes 57

14 Jars and Roots 60

15 Starred Paragraphs 64

IV Explanations 68

16 What It Could Be 69

17 Attempting a Reading 73

18 A Defensible Reading 78

19 Judging the Next Solution 81

A Data and Methods 84

B Folio Concordance 87

C Statistical Tables 90

D How This Book Was Made 93

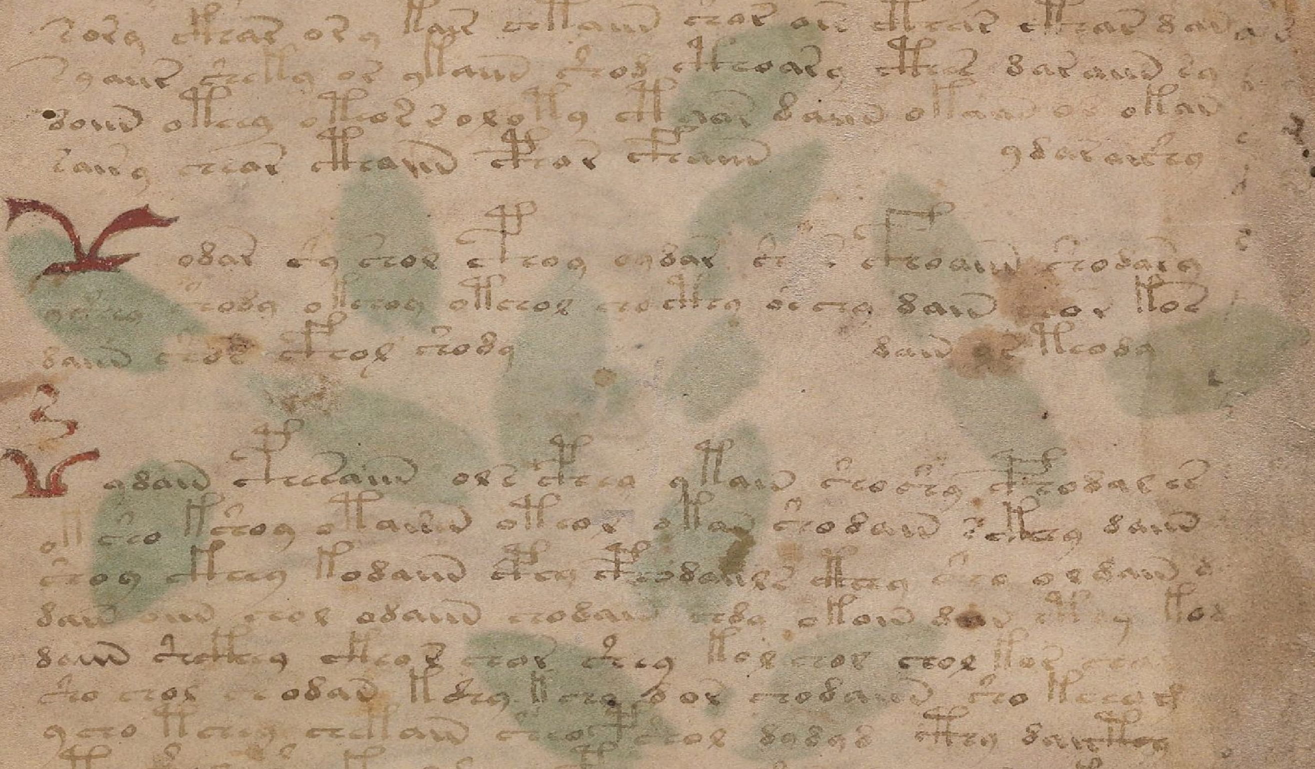
Bibliography 95

Prologue: the first page

FOLIO I recto is no title, no decorated initial, no owner's crest. The top lefts-

ing start. There is the page you see when you opened the book, and it is not a pr corner is scuffed where Wilfried Voynich, not long after he bought it, tried to bring up an erased signature with a chemical reagent and made it worse. Four paragraphs of text fill the page, separated by small gaps. The first begins with a tall, looping character that reaches up above the line like a gallows; the second and third begin with thick, odd marks that occur nowhere else in the book; the last begins with a gallows again. Each paragraph ends with a short word or phrase pushed to the right margin, the way a quotation might end with the name of its author. In that same margin, in a different and later hand someone has written a column of ordinary Latin letters, *a, b, c*, as if trying a key that did not fit. Everything about the layout says “book.” The margins are ruled by eye but kept. The lines are level. The words are separated by spaces of the kind a scribe leaves without thinking. The ink is the brown of iron-gall ink that has aged on calfskin. If you have handled fifteenth-century manuscripts, your hands know what this is before your eyes do. Then you look at a word. The first one on the page is transcribed, in the alphabet scholars use for this script, as **fachys**. Nobody knows how to say it. Nobody knows whether it is a word at all, in the sense of a unit of sound that means something, or a group of symbols standings for something else, or nothing. The second word is **ykal**; the third is **ar**, the fourth is **ataiin**, and **ataiin** is the first hint that the text has rules, because that ending, -iin, will turn out to close more than four thousand of the words in the book, and to appear anywhere but at the end only about a hundred times.

This book is about those rules, and about everything else that can be established about the object without being able to read it. It is written in the belief that the right response to an unsolved problem is not a solution but an inventory of what is known, how well, and from what evidence. The Voynich Manuscript has attracted more proposed solutions than any other text of its size, and nearly all of them have failed in the same way, by starting from an answer and working backwards. We start from the page.



Beinecke MS 408, f. 1r. The opening page: four paragraphs, each ending in a short right-justified word; the erased signature at top left; a later alphabet in the right mar-

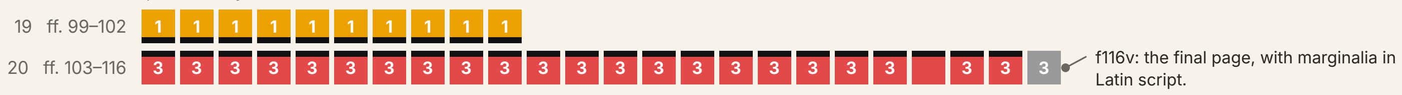
Part
I

The Object

1

A Book Nobody Can Read

THE first thing to say about the Voynich Manuscript is how



Source: Zandbergen–Landini transliteration ZL3b (May 2025), page variables; scribes after L. F. Davis, Manuscript Studies 5 (2020). Chart: the authors.

Figure 1.1 The manuscript as a grid.

ordinary it looks. It is a small, fat book of calfskin, about the size of a modern hardback novel, in a limp vellum cover that was put on it a few centuries after it was made. Open it anywhere in the first half and you find what you would find in a hundred other fifteenth-century books a plant, drawn in brown ink and washed with green and a muddy red, with a block of text wrapped around it. The text runs left to right. The lines are straight, the margins are tidy, the hand is fluent. It is only when you try to read a word that the ordinariness ends.

The script has never been read. Not in 1639, when a Prague alchemist named Georg Baresch wrote to the most famous scholar in Europe asking for help with

ls; not in 1912, when the book dealer Wilfrid Voynich bought it from a Jesuit college outside Rome and began showing it to anyone who would look; not in the century since during which it has been attacked by the cryptanalysts of two world wars, by linguists, botanists, astronomers and historians of medicine, and by a very large number of people with a theory and an afternoon. Yale University, which has owned it since 1969, catalogues it as *Cipher manuscript* and gives its language as “Undetermined” (Beinecke Rare Book and Manuscript Library, 2020).

This book is about what can actually be known about it. That turns out to be a surprising amount, provided one is strict about the difference between what the object shows, what the text measurably does, and what one would like either of them to mean. The calfskin can be dated: it was prepared between about 1404 and 1438. The quires can be counted and their original order partly recovered. The script can be transliterated, and once it is, the text can be measured in the same way any text can—how often each word recurs, how predictable each character is from the one before, how a line begins and how it ends—and

whose measurements place firm limits on what it could be. None of this reads the book. All of it narrows the question.

WHAT THIS BOOK DOES NOT CLAIM

It is worth being explicit, because the Voynich literature is full of books that are not. We do not offer a decipherment. We do not identify the language, the author, the place of writing, or the purpose. Where we describe a plant as a plant or a batch as a batch we mean only that this is what the drawing shows; we do not claim to know what it is for. We use the word “word” for a group of symbols between two spaces, and “letter” or “glyph” for a symbol; without implying that either behaves like its namesake in a language.

Every number in the chapters that follow was computed from a published machine-readable transcription of the manuscript, the Zandbergen–Landini transliteration, using short programs that are printed in Appendix A and that anyone can run (Zandbergen, 1996–2026). Every statement about the physical object rests on the published conservation studies, the radiocarbon result, and the digitised folios, which Yale has placed online at full resolution and which are reproduced throughout this book (Clemens, 2016; Hodgins, 2016; Zyats et al., 2016). Where we rely on someone else's observation we say so. Where we are guessing we say that too.

WHAT IT DOES CLAIM

What we offer is a reading in the older sense of the word: a careful look at what is there. The manuscript has a structure at every level, from the stroke to the quire, and that structure is evidence. Some of it was noticed long ago.

Prescott Currier, a US Navy cryptanalyst, saw in the 1970s that the text comes in two statistical dialects and that the line of writing behaves as a unit (Currier, 1976). John Tiltman, who had broken Japanese military codes, saw in the 1950s that the words are built from a small number of parts in fixed order (Tilman, 1967). Lisa Fagin Davis, a paleographer, established in 2020 that five people wrote it, and in 2026, with Colin Layfield, that its leaves were almost certainly bound in the wrong order by someone who could not read them (Davis, 2020; Layfield and Davis, 2026). Claire Bowern and Luke

Lindemann, linguist, showed in 2021 that at the level of characters the text is more predictable than any of the 250 languages they compared it with, and at the level of words and lines looks like ordinary language (Bowern and Lindemann, 2021). We have redone the measurable parts of that work from the current transcription, added several tests of our own, and arranged the whole into an argument. The argument runs as follows.

The object is a genuine early-fifteenth-century book, made with care by a small group of people working to a plan (Part I). The script is a real writing system in the sense that it has consistent rules, and those rules are unlike those of any known alphabet: nearly every symbol has a fixed position within the word, words are built from a few slots, the line is a unit, and the text has far less character-level variety than any European language written plainly (Part II). The pictures are of recognizable medieval genres—a herbal, a zodiac calendar, a bathing or medical sequence, a cosmological diagram, a pharmacy—drawn by people who knew the conventions but did not follow any surviving model closely (Part III). Taken together, these facts exclude several popular explanations outright and leave a small number standing of which the strongest is that the text encodes something, by a method that compresses the alphabet rather than expanding it, and that no one has yet reversed (Part IV). That is not a solution. It is a map of where a solution would have to be.

HOW TO USE THIS BOOK

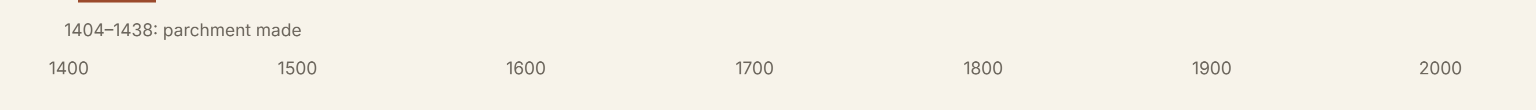
The chapters can be read in order or dipped into. Readers who want the pictures first should start with Part III, which can be read without Part II. Readers who want to know why every proposed decipherment so far has failed should go to chapter 18, which sets out the tests any solution has to pass, and which, we hope, will save some people a great deal of time.

Throughout, Voynichese words are given in the EVA transliteration (Landini and Zandbergen, 1998), in which the first word of the manuscript is **fachys**. EVA is a convention for writing the glyphs with Latin letters so that they can be typed and counted; the letters say nothing about pronunciation. Folio references follow Yale's foliation: f. 1r is the front of the first leaf, f. 1v its back. The large fold-out leaves have panels, numbered f. 70v1, f. 70v2 and so on, following the convention used by Zandbergen.

2

Skin, Ink and Pigment

BEFORE anyone asking what ink is, it is worth ask what the book says,



Sources: Hodgins (2014) for the dating; Zandbergen, voynich.nu, and Clemens (ed.) 2016 for the provenance.

Figure 2.1 What is documented about the book's history.

made of, because the materials answer two questions that the text cannot: when, and whether it is what it appears to be.

THE LEAVES

The text block consists of 102 leaves of parchment, with one modern paper leaf

Skin, Ink and

at the back, foliated 1 to 116 with gaps (Beinecke Rare Book and Manuscript Library, 2020) A standard leaf measures about 225 by 160 millimetres, which makes a standard sheet, folded once down the middle to give two leaves, about 320 by 230. The leaves are gathered into eighteen surviving quires, numbered i to 20; quires 16 and 18 are missing entirely, and six further sheets are missing from the centres of quires 8 and 20 and from two other places, so that fourteen of the original 116 folios are gone (Zandbergen, 1996–2026). Most quires are the ordinary kind, four sheets nested to give eight leaves, but the book is unusual in the number of leaves that are wider than standard and fold out, some to three panels, some to four, one to six, and one sheet, the famous “rosettes” page, to roughly 45 by 45 centimetres with an extra horizontal fold. Fold-outs are not rare in medieval books, but this many in a book this small is.

The parchment is calfskin. That was established in 2014 by protein analysis at the University of York, which can tell the species of animal from a sample of collagen too small to see (Zandbergen, 1996–2026). “Vellum,” strictly, means calf, so for once the commoto name is right.

THE DATE

In 2009 Greg Hodgins of the University of Arizona's accelerator mass spectrometry laboratory took four small samples from four different leaves of the manuscript and dated them by radiocarbon. The four results agree with one another, and their combined calibrated range is 1404 to 1438 at 95 per cent confidence (Hodgins, 2016). That is the date at which the calves died and the skins were made into parchment. Parchment was expensive and was not usually stockpiled for decades, so the writing is most likely from the same generation; but the date is a floor, not a ceiling. What the radiocarbon result rules out is any theory that places the book's manufacture in the sixteenth century or later, including every version of the story in which the Elizabethan magus John Dee, or Wilfrid Voynich himself, made it. The date is consistent with everything else. The quire numbers, written in the lower corners of the last leaf of each gathering, use a form of numeral and an abbreviation for the Latin ordinal (*primus*, *2us*) that belong to the fifteenth century (Davis, 2025; Zandbergen, 1996–2026). The drawings of the zodiac, especially the crossbow-carrying Sagittarius, match conventions found in the German manuscripts of roughly 1400 to 1500 (Zandbergen, 1996–2026).

The women's headdresses and the shape of the few recognisable objects in the drawings point the same way.

THE INK AND THE PAINT

The text and the outlines of the drawings are in an iron-gall ink, the ordinary writing ink of medieval Europe, made from oak galls and iron salts. Analysis of the ink and pigments was carried out for Yale by McCrone Associates in 2009 and summarised in the 2016 Yale volume on the manuscript (Zyats et al.,

Skin, Ink and

2016). The blue is azurite, a copper carbonate mineral; the greens are copper-based; the red-brown is an iron oxide earth. These are the cheapest and commonest pigments of the period. There is no gold, no expensive lake, no ultramarine, nothing a wealthy patron would have paid for. A yellow wash and some brown washes are also present. The ink and the paint are consistent with the date and with each other, and nothing about them is anachronistic.

Two further observations about the paint matter for what follows. The first is that the paint was applied after the ink, in at least two campaigns, one careful and one much less so; several of the plates have been coloured by someone who did not keep inside the lines, and in a few places paint has been transferred to the facing page while still wet (Zandbergen, 1996–2026). The second follows from the first: because some of those wet-paint transfers main the page that currently faces them, the clude coloring was done after the leaves had reached their present order. We return to this in chapter 4, because

itch bears on when the book was disordered and by whom.

THE COVER AND THE SEWING

The cover is a limp vellum case of the late eighteenth or nineteenth-century, probably put on in Rome by the Jesuits, who owned the book by then (Beinecke Rare Book and Manuscript Library, 2020; Davis, 2025). It is not original, and the Yale catalogue notes that it was once stiffened and had paper paste-downs, both since removed. Underneath, the quires are sewn onto thongs. The leather thongs visible now were added in a repair in the 1960s and glued over older ones of some kind of fibre which may be the original fifteenth-century sewing supports (Zandbergen, 1996–2026). The sewing holes in the folds of the leaves are worn, and in at least one quire there is a second set of holes that does not match the present binding, which is one of the pieces of evidence that the book has been taken apart and put back together.

IS IT GENUINE?

The question is still asked, so it is worth answering directly. Every physical test applied to the manuscript has come back consistent with an early-fifteenth-century European book: the parchment, its date, the ink, the pigments, the quire signatures, the wear, the sewing. A forger working before 1912 would have needed to acquire well over a hundred leaves of blank fifteenth-century calfskin, including several of the oversize sheets, and would have needed to anticipate protein analysis and accelerator radiocarbon dating. The alternative, that Voynich or a predecessor found a blank fifteenth-century book and filled it, founders on the same point: blank medieval books of this size and construction do not survive, because blank parchment was reused. What the materials cannot tell us is whether the text means anything. A real

fifteenth-century book can contain nonsense as easily as sense. That question belongs to the text, and to Part II.

The limp vellum covers, added in the eighteenth or nineteenth Century. Beinecke Rare Book & Manuscript Library, Yale University.

Inside the front cover: traces of the removed paper paste-down. Beinecke Rare Book & Manuscript Library, Yale University.

3

Where It Has Been

A Book’s history is usually written in it: owners sign their names, librarians add shelf-marks, readers scribble. This book has very little of that, and the little it has is faint. Its documented history begins nearly two centuries after it was made, in Prague, and for the century before that there is nothing at all. What follows is the record as it stands, with the gaps left open (Clemens, 2016; Zandbergen, 1996, 2016, 2016).

PRAGUE

At the foot of the first page, invisible in ordinary light and only partly legible under ultraviolet, is a signature: *Jacobj à Tepenecé*, with what appears to be a shelf-mark beneath it. Jakub Hořčický, a pharmacist and distiller who had made money and friends at the court of the Emperor Rudolf II, was ennobled in 1608 with the title "de Tepenec," and died in 1622. The signature therefore dates from between those years, and is the earliest certain fact about the book's ownership. Someone erased it, probably not long after, Wilfrid Voynich's attempt to revive it with a chemical reagent in the 1910s is the reason the corner of the page looks as it does (Davis, 2025).

The next witness is a letter. In 1639 Georg Baresch, an alchemist living in Prague, wrote to the Jesuit polymath Athanasius Kircher in Rome, who had just published a book claiming to read Egyptian hieroglyphs, and sent him a copy of some pages of a manuscript he could not read, hoping Kircher could (Zandbergen, 1996–2026). Baresch describes the book: it is full of plants unknown to Germans, pictures of stars, and “chemical secrets.” Kircher’s reply does not survive, but the book stayed with Baresch until he died, when it passed to his friend Johannes Marcus Marci physician and rector of the university in Prague. Marci sent the book itself to Kircher in 1665 or 1666, with a covering letter that is now tucked inside its front cover. The letter is the source of nearly everything that has been said about the book's early history. Marci tells Kircher that a former owner, Rudolf II, had paid six hundred ducats for it, and that the Emperor had believed its author to be the thirteenth-century English friar Roger Bacon. Marci is careful to attribute the Bacon story to a third party, Raphael Mnishovsky, a courtier who had died in 1644, and to say that he himself suspended judgement. No record of the purchase has been found in the imperial accounts, and Mnishovsky is not a reliable witness; but Tepenec's signature puts the book in Rudolf's Prague, and the sum, though large, is not impossible for a book believed to be by Bacon (Zanberger, 2016).

Whether Rudolf had the book from John Dee, the English magus who was in Bohemia in the 1580s, is a story with no evidence behind it beyond the fact that Dee owned Bacon manuscripts and was in the right place. The older claim that the folio numbers are in Dee's hand is not accepted by palaeographers who have examined them (Davis, 2025).

ROME

Kircher's library passed, at his death in 1680, to the Jesuit college in Rome, the Collegio Romano. The book is not mentioned in any sur-

Where It Has

viving catalogue there, which is not surprising, it was a curiosity, not a text. When the Italian state confiscated church property in the 1870s, the Jesuits moved their more valuable books into the private libraries of individuals to keep them out of the state's hands. The manuscript went into the collection attributed to Petrus Beckx, the Superior General of the order, and from there to the Villa Mondragone at Frascati, outside Rome,

which the Jesuits used as a college (Hunt, 2016).

LONDON, NEW YORK, NEW HAVEN

In 1912 the Jesuits, short of money, quietly sold a group of manuscripts from the

Villa Mondragone to Wilfrid Voynich, a Polish-born rare-book dealer with shops in London and, later, New York. Voynich bought about thirty; this was one of them.

The sale was supposed to be confidential, and Voynich obscured the book's origin for the rest of his life, which did nothing to help later researchers (Hunt, 2016).

Voynich was not a quiet owner. He showed the book widely, had photostats made and circulated them to anyone who might read it, and promoted the Bacon attribution as far as it would go, which was quite far: in 1921 a professor of philosophy at the University of

Pennsylvania, William Romaine Newbold, announced a decipherment based on microscopic marks within the strokes, in which Bacon had described the Andromeda nebula and the structure of cells. The marks turned out to be the cracking of the ink. Newbold's readings was demolished by John Manly in 1931, and the episode set the pattern for the next century: a confident solution, wide publicity, and a collapse on inspection (D'Imperio, 1978).

died in 1930. The book passed to his widow, Ethel Lilian Voynich, the novelist, and on her death in 1960 to her companion Anne Nill, who sold it

Voynich

in 1961 to the New York dealer Hans P. Kraus for $24,500. Kraus priced it at $160,000, found no buyer, and in 1969 gave it to Yale University's Beinecke Rare Book and Manuscript Library, where it was catalogued as MS 408 (Clemens, 2016). It has been there since, photographed in full several times, and since 2020 the complete digital facsimile has been free to anyone with a web connection (Beinecke Rare Book and Manuscript Library, 2020).

WHAT THE RECORD SUPPORTS

Set out

Where It Has

this way, the provenance is thin but coherent. The book was in Prague by about 1610, at the court where Rudolf II, collected exactly this kind of thing. It passed through the hands of three men who tried to read it and could not. It spent two and a half centuries in Jesuit libraries, unread. It was bought by a dealer who tried to sell it as the work of a famous Englishman, failed, and left it to his library. At no point in that chain does anyone claim to understand it, and no one in the chair had the means to make it.

Everything before 1610 is inference from the object itself: the parchment date, the style of the drawings, the fifteenth-century quire numbers. On those grounds alone the book was made somewhere in central Europe, most likely in the German-speaking lands or northern Italy, in the first half of the fifteenth century. Where exactly, and by whom, the record does not say.

f. 1r, detail. The two paragraph-opening marks in red ink, unlike anything else in the manuscript, at the start of the second and third paragraphs. Beinecke Rare Book

& Manuscript Library, Yale University.

Where It Has

4

In What Order?

A Reader assumes that a book's pages are in the order its maker intended. For this book the assumption is probably wrong, and the evidence that it is wrong has been accumulating for twenty years. It reached a new stage in October 2026, when Colin Layfield and Lisa Fagin Davis published a statistical study arguing that the leaves were never meant to be bound as they are (Layfield and Davis, 2026). This chapter sets out what is known about the order of the book, which is more than one might expect, and what it means for reading it.

THREE LAYERS OF NUMBERING

There are three systems of numbers on the leaves, and they were added at different times by different people.

The quire marks come first. On the last leaf of each gathering, in the lower corner, someone wrote the number of the gathering as a Latin ordinal: *primus*, 2*us*, 3*us*, and so on to 20, with the "-*us*" written as a superscript 9, in the way fifteenth-century scribes abbreviated the ending. The numerals are in a hand and a style that palaeographers date to the fifteenth century, within a century of the writing of the book (Davis, 2025; Zandbergen, 1996–2026). Quire marks are added when a book is bound, so the book was bound in something like its present arrangement of gatherings while the century was still young. Quire marks 16 and 18 are missing because those gatherings are missing; marks 19 and 20 differ slightly in form from the rest, which Nick Pelling has taken to suggest more than one numberer (Pelling, 2006). The folio numbers come second. They are written in the upper right corner of each leaf in arabic numerals, in a hand that Davis places in the seventeenth century and that Yale's catalogue places in the sixteenth (Beinecke Rare Book and Manuscript Library, 2020; Davis, 2025). They run from 1 to 116 without resetting, which tells us two things: that the gatherings were in their present order when they were numbered, and that all 116 leaves were present. The fourteen leaves now missing were lost after the numbering, and in at least one case the loss is visible; a stub at folio 12 shows where a leaf was cut out (Davis, 2025). The foldout leaves were numbered folded, so that their panels do not always carry the number where one would expect it, and the modern convention for naming the panels, f.70v1, f.70v2 and so on, was invented in the twentieth century. The third system is modern: the panel names, and Yale's catalogue numbering of the digital images

THE EVIDENCE THAT THE ORDER IS WRONG

What makes anyone doubt that the present order is the original one? René Zandbergen has collected the arguments over many years (Zandbergen, 1996–2026), and Davis has added physical evidence (Davis, 2025). The principal points are there.

Subject matter. The herbal pages are not all together. After fifty-six leaves of plants come the astronomical, zodiac, and bathing sections, and then, in quires 15 and 17, ten more leaves of plants that are not visibly different from the first fifty-six. The pharmaceutical pages are split between quires 15 and 19, and the style of the drawn containers suggests that the two leaves in quire 19 should come before the one in quire 15.4

Scribes. In quires 4 to 7 of the herbal section, the work of Davis's Scribe 1 alternates with that of Scribe 2, sheet by sheet: the outer sheet of quire (fo-

In What Orlios 25 and 32) is by Scribe 1, the next (26 and 31) by Scribe 2, the two inner sheets by Scribe 1 again. The same alternation produces the scattered language-B pages in figure 1.1. Davis, who has examined a great many medieval manuscripts, says she has never seen a book organised this way, and the simple explanation is that sheets by the two scribes were once grouped and have since been shuffled.

A picture across the fold. On the back of folio 78 and the front of folio 81, in the bathing sections, water drawn flowing from a vessel on the one page continues onto the other. The two pages form one design, which can only have been seen when the sheet they belong to lay open at the centre of its gathering. It does not; another sheet is bound inside it (Zandbergen, 1996–2026).

A stain. A large water stain runs through the upper margins of the first several dozen leaves. If the leaves had been in their present order when the water struck, the stains would shrink or grow steadily from leaf to leaf. It does not on folio 32v it is smaller

than on the facing 33r. Davis concludes that the book got wet, was taken apart to dry, and was put back together in a different order, all blefore the seventeenth-century folio numbers, which sit on top of the stains, were added (Davis, 2025).

Sewing holes. In quire 9, a single large fold-out sheet, there is a set of sewing holes in a fold other than the one now in the spine. If the sheet had been bound at that fold, its quire mark would sit where quire marks normally sit. The sheet was refolded and resewn (Davis, 2025; Zandbergen, 1996–2026).

Paint. As noted in chapter 2, some of the clumsier paint has offset onto facing pages, and the offsets match the pages that face them now. The careless painting therefore happened after the reordering. Whoever coloured those plants did not know what order the book should be in, which means, as Zandberger points out, that the colours are not a reliable guide to the plants (Zandberger, 1996–

2026).

THE SINGULION HYPOTHESIS

Layfield and Davis's 2026 paper adds a statistical argument and a new proposal (Layfield and Davis, 2026).

They applied latent semantic analysis, a method for measuring the similarity of texts by their vocabulary, to blocks of the transcribed text at the level of pages, leaves, sheets and quires, and ran the same tests on two ordinary medieval books, the Aberdeen Bestiary and a copy of the *Speculum humanae salvationis*, as controls. In the controls, adjacent pages resemble one another in the way one expects of continuous text. In the Voynich manuscript, some adjacent pages do not, and some pages far apart do. Their proposal is that the book was conceived not as a set of nested gatherings but as a sequence of single folded sheets—“singulions” in the authors’ term—each a unit in itself, to be read one after another. A binder who nested those sheets into quires without being able to read them would produce exactly the pattern observed: scribes alternating sheet by sheet, a picture split across a fold, a text whose near neighbours are not always its nearest relatives. The authors are explicit that this re-orders the problem rather than solving it. It does not produce a translation, and the original sequence of the sheets has not yet been recovered, they propose to use the waterstain, whose shape should be continuous in the true order, as a guide.

The paper is three days old as this book goes to press and has not yet been examined by other researchers. Some of its conclusions rest on measurements of the text that depend, as all such measurements do, on the transcription and on choices about what counts as a unit. The physical arguments do not depend on those choices, and they were already strong.

WHAT FOLLOWS FOR A READER

Three things follow.

First, any analysis of the text that treats page order as meaningful—

looking for a narrative, for a table of contents, for an index at the back, for a first chapter—is standing on sand. Analyses that treat the page, or better the sheet, as the unit are on firmer ground, and that is how the rest of this book proceeds. Second, the sections are real but their boundaries are approximate. A leaf of plants in quire 17 is a herbal leaf; where it originally stood, we do not know.

Third, and more hopefully, the disorder is a kind of evidence. If the sheets can be put back in order by their physical traces, the result can be checked against the text:

a reconstruction that also makes the text's vocabulary flow continuously from sheet to sheet would be a strong sign that the text means something and that the measurements are tracking it. That test has not been run. It is one of several we recommend in chapter 19.

f. 78v: water flows from the vessel at the right edge... Beinecke Rare Book & Manuscript Library, Yale University.

and continues on f. 81r. The two pages were drawn as one opening; another sheet is now bound between them. Beinecke Rare Book & Manuscript Library, Yale University.

Part

II

The Script

5

Writing It Down

NOTHING can be counted until it has been written down in a form that can be counted, and for this manuscript that step is neither obvious nor finished. Every statistic in this book, and every statistic in the literature, rests on a transliteration, a decision about what the distinct symbols are, and a rendering of each page as a string of Latin letters standing for them. The decisions are not innocent. Some of the most-quoted “facts” about the text are facts about the transliteration.

Figure 5.1 Two readers, one page.

Figure 5.2 The twenty commonest words, by language.

WHAT THE SYMBOLS ARE

Look at a line of the text and the first impression is of an alphabet of modest size, perhaps twenty to thirty characters, written with a few basic strokes. Several characters resemble Latin letters or the abbreviation marks of Latin manuscripts: a form like *n*, a form like *h*, *o*, a form like the numeral *9*, used by scribes for *un*, a form like 4, and a looped form that medieval scribes used for -*is* or -*ris*. Four tall characters with loops on top—the “gallows”—have no close parallel. A pair of forms like m joined by a bar across the top, which scholars call benches, can also be written with a gallows letter planted in the middle, producing composite forms. The difficulty is deciding where one symbol ends and the next begins. Is the sequence that EVA writes **ch** one letter or two? The strokes are joined, and it behaves in the text as a unit. Is **iii** three strokes of one letter, like the minims of *m*, or a run of three? Is the gallows-in-a-bench one character or two superimposed? Davis, using the Archetype digital palaeography platform to compare every instance,

concluded in 2020 that the inventory is small and that the scribes wrote the same forms the same way, which is what one would expect of a practised script and not of an improvised one (Davis, 2020).

FOUR ALPHABETS

Over the past seventy years four main transliteration alphabets have been used (Zandbergen, 1996–2026). The First Study Group under William Friedman produced one in the 1940s and 1950s. Prescott Currier produced his own in the 1970s, using capital letters and digits (*4OFCC89* for the word EVA written **qokeedy**). In the 1990s Gabriel Landini and René Zandbergen designed the European Voynich Alphabet, EVA, which uses lowercase Latin letters chosen so that the transliterated text is pronounceable and easy to type; it has been the standard for machine-readable work since. Glen Claston's v101 alphabet, from the 2000s, splits more finely, distinguishing variant forms that EVA merges.

The alphabets differ in what they treat as a unit. EVA writes the commonest bench as two letters, **ch**, and the three-stroke ending as **in**; Currier wrote each as a single symbol. Because EVA writes some single glyphs as two or three letters, statistics computed on EVA text are not quite statistics of the script. Bowern and Lindemann make the point carefully: EVA “probably underdifferentiates” some glyphs and overdifferentiates others, and the choice affects measurements such as entropy

(Bowern and Lindemann, 2021). Throughout this book we therefore give results on the EVA letters and, where it matters, on a "glyph-unit" version in which the benches **ch**, **sh** and the four gallows-benches are counted as single symbols. The conclusions do not change; the numbers move a little.

A transliteration alphabet is only useful once someone has applied it to all 226 pages and foldout panels and that has been done several times, by different people, with different results. The Zandbergen–Landini transliteration, begun in 1999 and revised and continuously since, is the most complete it records every piece of writing on every page, including labels, circular text and marginalia, with the position of each, and a set of variables for each page recording its quire, its illustration type, its Currier language and its scribe. We use its version of May 2025, ZLib, throughout (Zandbergen, 1996–2026) Takeshi Takahashi's transliteration of the late 1990s, made independently, is the other full one in EVA. The two can be compared line by line, and the comparison is the best available measure of how much of the "text" is actually settled. On the 5,214 lines both transcribers covered, the two agree exactly on 47 per cent of lines. On a word-by-word alignment they agree on 75 per cent of words. The disagreements are mostly about spaces—whether a gap is wide enough to count as a word break—and about a handful of ambiguous letterters, forms, **a** against **o**, **r** against **s**, the number of strokes in an **i**-run. Both transliterations mark their own uncertain spaces and uncertain let-

and, we have followed the Zandbergen–Landini markings, treating uncertain spaces as spaces and dropping the 140 words that contain an unreadable character.

The practical consequence is that any claim resting on the exact spelling of an individual word is weaker than it looks. Claims resting on thousands of words, where the transcribers' disagreements wash out, hold up, and those are the claims we make.

THE TEXT AS COUNTED

With those cautions, here is

the text by the numbers. The ZL3b transliteration contains 5,225 separate items of writing: 3,885 lines of ordinary paragraph text, 299 labels

on the zodiac pages, 194 labels on plant parts in the pharmaceutical section, 76 labels of stars, 63 of the small female figures in the bathing section, and the rest circular and radial text in diagrams and miscellaneous labels. Together they hold 37,691 words. The ordinary running text, which is the basis for most of what follows, contains 34,180 words after the unreadable ones are removed, of 7,466 distinct spellings.

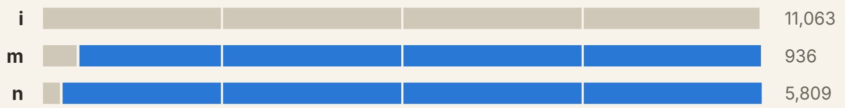
The 100 commonest spellings account for 44 per cent of the running text and the 500 commonest for two thirds; 5,266 of the 7,466 distinct words occur only once. The commonest word, **daiin**, occurs 764 times, 2.2 per cent of the running text. The next are **ol**, **chedy**, **aiin**, **shedy**, **chol**, **or**, **chey**, **ar** and **qo-**

keedy Figure 5.2 gives the twenty commonest and shows at once the single most important division in the text, the words that dominate the first half of the list, **daiin**, **chol**, **chor**, nearly vanish in some sections, and the words that replace them there, **chedy**, **shedy**, **qokeedy**, are almost unknown elsewhere. That division is Currier's two "languages," and it is the subject of chapter 8.

6

Every Letter Knows Its Place

IN 1951 John Tiltman, a British cryptanalyst who mid-word.



y ends 85 % of its words; n and m end 97 % and 94 %, and m clusters at the ends of lines.

Source: ZL3b transliteration, running text only; EVA alphabet with ch, sh, cth, ckh, cph, cfh counted as single symbols. Counts: the authors.

Figure 6.1 Where each symbol sits in the word.

had spent the war on Japanese military ciphers, was shown photostats of the manuscript by William Friedman and set to work on the words. What he found, and reported with some puzzlement, was that they were built like compounds from a kit. Each word seemed to consist of a beginning drawn from one small set of elements, a middle from another, and an ending from a third, and most letters belonged to only one of the three sets (Tiltman, 1967). He could not say what the parts meant. But the observation has been confirmed by everyone who has looked since, and it is the single most important structural fact about the script.

POSITION

Figure 6.1 shows the fact directly. For each of the twenty-two commonest symbols it shows how its occurrences divide between the start of a

Figure 6.2 Word length, against five reference texts.

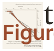


Figure 6.3 Rank against frequency.

word, the inside, and the end. Three groups stand out.

At the top is **q**. It occurs 5,345 times in the running text, which makes it the twelfth commonest symbol, and 99 per cent of the time it is the first letter of its word. In 97 per cent of those cases the second letter is **o**. No letter of any European alphabet behaves like this; the nearest analogy is a letter that is really a prefix, or a mark that is written before a word rather than as part of it. Below **q** come the benches, **sh** and **ch**, and the gallows-benches **cth**, **ckh**, **cph**, **cfh**, which open a word two or three times out of four and never closes one.

At the bottom are the closers. **n** ends 97 per cent of its words and **m** 94 per cent; neither even begins one. **y** ends 85 per cent of its words and **r** three quarters. **l** is split, and is the one symbol that is at home anywhere.

In the middle

are **e**, **i** and **a**, which are never first and never last. They occur in words, runs **ee**, **eee**, **ii**, **iii**. The symbol **o** is the commonest in the text, and is mostly a middle letter, but it begins a fifth of all words, usually before a gallows.

THE GALLOWS

The four tall letters **k**, **t**,

p, **f** occur in just over half of all words, and almost never twice in the same word. 1.5 per cent of words contain two. They sit in a characteristic place, after the beginning and before the middle: **okaiin**, **qokeedy**, **otedy**, **choty**. Of the four, **k** and **t** are common, **p** and **f** rare, and

p and **f** have a strong preference for the first line of a paragraph, which is the subject of the next chapter. Whatever the gallows are, a word has room for one of them, in one slot.

A GRAMMAR OF SLOTS

Pittman's kit can be made precise. Jorge Stolfi, in the late 1990s, proposed a "core–mantle–

crust” model in which every word has a core (typically a gallows or a bench), wrapped in a mantle of **e** and **o** and **a** letters, wrapped in a crust of the opening and closing letters (Stolfi, 1997–2000). There are several such models, and the differences between them matter less than what they share.

To make the point without choosing between them, we wrote the simplest version we could: an opening slot (one of **qo**, **o**, **d**, **s**, **ch**, **sh**, **y**, **a**, **l**, **k** or nothing), a gallows slot (one of the eight gallows forms or nothing), a middle slot (one of **ch**, **sh**, a run of **e**, **eo**,

o, **a** or nothing, an optional single **d**, **l**, **r** or **s**, and a closing slot (one of **y**, **dy**, **ain**, **aiin**, **an**, **ar**, **al**, **or**, **ol**, **am**, **ey**, **eey**, **edy**, **eedy**, a single closing letter, or nothing). That template, which fits on three lines, matches 78 per cent of all word tokens in the running text, and a third of the distinct spellings.

What it fails to match is instructive. The commonest unmatched words are **dair**, **air**, **cheky**, **chckhy**, **olkeedy**, **olkeey**, **choty**, **olchedy**, **olkain**. Most of these are two complete words run together: **olkeedy** is **ol** followed by **keedy**, **olchedy** is **ol** and **chedy**, **cheky** is **che** and **ky**. The frequent short words **ol** in particular attach to the front of other words so often that it behaves like a clitic. A more generous template that allows two cores covers over 90 per cent of tokens; the residue is mostly transcription noise and the long strings in labels.

LENGTH

A conse-

Every Letter

quence of a building words from a fixed number of slots is that they come out at nearly the same length. Figure 6.2 shows the distribution. Voynichese words have a mean length of 5.1 EVA letters and a standard deviation of 1.9, a quarter of all words are exactly five letters long. The diserence texts, sampled to the same size, have broader distributions, and Latin's is much broader (standard deviation 2.7). What is striking is less the

mean, which depends on the transliteration, than the shape: a tight, nearly symmetrical peak, with very few one- and two-letter words and a tail that falls off fast. Natural languages have a long tail of long words and a fat head of short function words. Voynichese has neither.

WHAT THIS DOES AND DOES NOT SHOW

These regularities establish that the text was produced by a procedure with rules, applied consistently across thirty-seven thousand words by five different scribes. That already excludes the lazier kinds of nonsense: a person scribbling freely does not keep q at the front of words 99 times in 100 for two hundred pages. They do not, on their own, tell us what the rules are rules for. A slot structure of this kind can arise in several ways. It is what a syllabary looks like when written out: consonant, vowel, consonant. It is what an abbreviation system looks like. It is what a cipher looks like when each letter of the plaintext is replaced by a group of symbols drawn from a fixed nable a so-called verbose cipher, in which the plaintext letters show through as the slots (Bowern and Lindemann, 2021). And it is what a well-made

generating procedure for meaningless text would produce, if its designer wanted the result to look like a language (Timm and Schinner, 2020).

Every Letter

Telling these apart requires the rest of the evidence, and chapter 16 takes it up.

Two things can be said now. The first is that the slogs are not arbitrary: they recur in exactly the places where a writing system would put them, and they interact with the line and the paragraph in ways that a generating procedure would have had to be designed to produce. The second is that the glyphs with the most fixed positions, **q** at the start and **n**, **m**, **y** at the end, are precisely the ones that are rare or absent on labels, where there is no line to be at the start or end of. That is the subject of chapter 7.

7

The Line and the Paragraph

IN line
is most
an writ-
ten ac-
text ci-
the dent.

Source: ZL3b, 4,034 lines of running text with three or more words. Counts: the authors.

Figure 7.1 How spelling changes with position in the line.

A scribe writes until he reached the margin and continues on the next line; the words do not know where the margin is. Prescott Currier noticed in the 1970s that in this manuscript they seem to (Currier, 1976). Words at the start of a line are spelled differently from words in the middle, words at the end are spelled differently again. The effect is large enough to see without statistics

once one knows to look, and it is one of the clearest signs that the layout on the page is part of whatever system produced the text.

Figure 7.2 Paragraph openings.

Figure 7.3 Labels against running text.

BEGINNINGS AND ENDS

Figure 7.1 shows the effect in the 4,034 lines of running text with at least three words. Take the first letter of each word. Among words that begin a line, 9 per cent start with **p** and 10 per cent with **t**; among words inside a line the figures are 0.5

and 2 per cent. Words beginning with **s**, **d** and **y** are also enriched at the line start, and words beginning with **ch**, then commonest opening

inside a line, falls from 21 per cent to 4. Now take the last letter. Among words that end a line, 16 per cent end in **m**; among words inside a line, under 1 per cent. Words ending in g, rare everywhere, occur almost only at line ends.

The words themselves change length. The first word of a line averages 5.45 letters, a middle word 5.08, the last word 4.88. The pattern is consistent across sections and scribes.

PARAGRAPHS

The paragraph

is a unit too. The Zandbergen–Landini transliteration marks the start of each paragraph; there are 736 of them. Of the 730 paragraph-opening lines with three or more words, 84 per cent begin with a word whose first letter is a gallows (**k**, **t**, **p**, **f**) or a gallows-bench. Of all other lines, 9 per cent do. Bowern, and Lindemann report the same proportion, 85 per cent, from their own count (Bowern and Lindemann, 2021). The rare gallows **p** and **f** are com-

concentrated there: 46 per cent of paragraph-opening lines begin with a **p**-word, against 1.3 per cent of other lines, and **f**, rare everywhere, is seventy times more likely at a paragraph opening than elsewhere.

Something similar happens at paragraph ends. On the first page, and on a few others (f. 8r, f. 25r, f. 42v, f. 46v), the final line of a paragraph is a single word or short phrase set off to the right, the kind of thing the transliteration calls a “title” and that Zanbergen has compared to the attribution at the end of a quotation. Elsewhere the last line of a paragraph is simply shorter, and ends, as we have seen, with a different set of letters.

WHAT KIND OF THING IS THIS?

There are two broad ways to read the line effects, and the literature has not settled between them.

One is that they are orthographic: a convention of writing, not of language. A scribe who writes **m** where he would otherwise write **in** when he reaches the end of a line, or who adds a flourish to the first letter of a page. A paragraph, is doing what medieval scribes did with Latin, where line-final abbreviations and enlarged initials are routine. On this view the words at line ends are the same words as elsewhere, in different dress. Bowern and Lindemann lean this way, calling the patterns “typographical” rather than linguistic” (Bowern and Lindemann, 2021). Some evidence supports it directly: the words that begin lines with **s** look like ordinary words that begin with **o** or **a** with an **s** in front—**saiin** for **aiin**, **sol** for **ol**—as though the **s** were a line-initial mark rather than a letter. The other reading is that the line is a unit of content: that each line is, or encodes, something complete, so that its first word and last word are a different kind of thing from its middle words, as the first and last words of a list, or a verse would be. On this view the effects are real structure in what-even the text is.

The two readings are not exclusive, and the data do not yet decide between them. What the data do establish is that the text was not produced by writing out a continuous stream and breaking it at the margin. Either the writer composed line by line, or he transformed what he wrote at the line boundaries, and either way the layout was in his hands as he wrote. This matters for the explanations considered in chapter 16: a cipher that enciphered a Latin text letter by letter would not know where the lines were, unless the plaintext had already been laid out in the same lines, which is possible but adds a step; a procedure for generating meaningless text would have to include a rule for line ends, which Timm and Schinner's does (Timm and Schinner, 2020).

LABELS

The strongest evidence that the line effects are about position comes from the words that have no line: the labels written next to drawings. The zodiac pages carry 345 label words, each beside a small figure; the pharmaceutical pages carry 202 beside drawings of plant parts and 36 beside containers; the bathing pages carry 64 beside figures and 55 beside tubes and vessels; 86 stand beside stabs. Labels differ from running text in several ways (figure 7.3). Fewer than 3 per cent begin with **q**, against 15.5 per cent of running-text words; over half begin with **o**, against 20 per cent. They are a little longer on average, and nearly all unique: of 281 distinct zodiac labels, only 37 occur on more than one zodiac page, and the container labels, with a mean length of seven letters, occur once each and almost never in the running text. Yet labels are made of the same parts. Half to four fifths of label words, depending on the type, occur as words in the running text as well; the zodiac labels **otaly**, **otal**, **okeey**, **oky** are ordinary vocabulary. The labels are, in short, words

from the same system, stripped of whatever **q** does and of the line-initial and line-final marks, which is what one would expect of words standing alone. Whether a label names the thing it stands beside is another question. In a herbal, the obvious expectation is that a label on a root is the name of the plant or its part; in a zodiac, that the label beside a figure is the name of a star or a degree. The expectation is reasonable and unproven. Attempts to use it as a key to read **otally** as a star name in some language—have so far produced readings

that do not survive transfer to the next label, let alone to the running text (Bowern and Lindemann, 2021). The labels are a resource for a decipherment, and we return to them in chapter 17. They are not, by themselves, one.

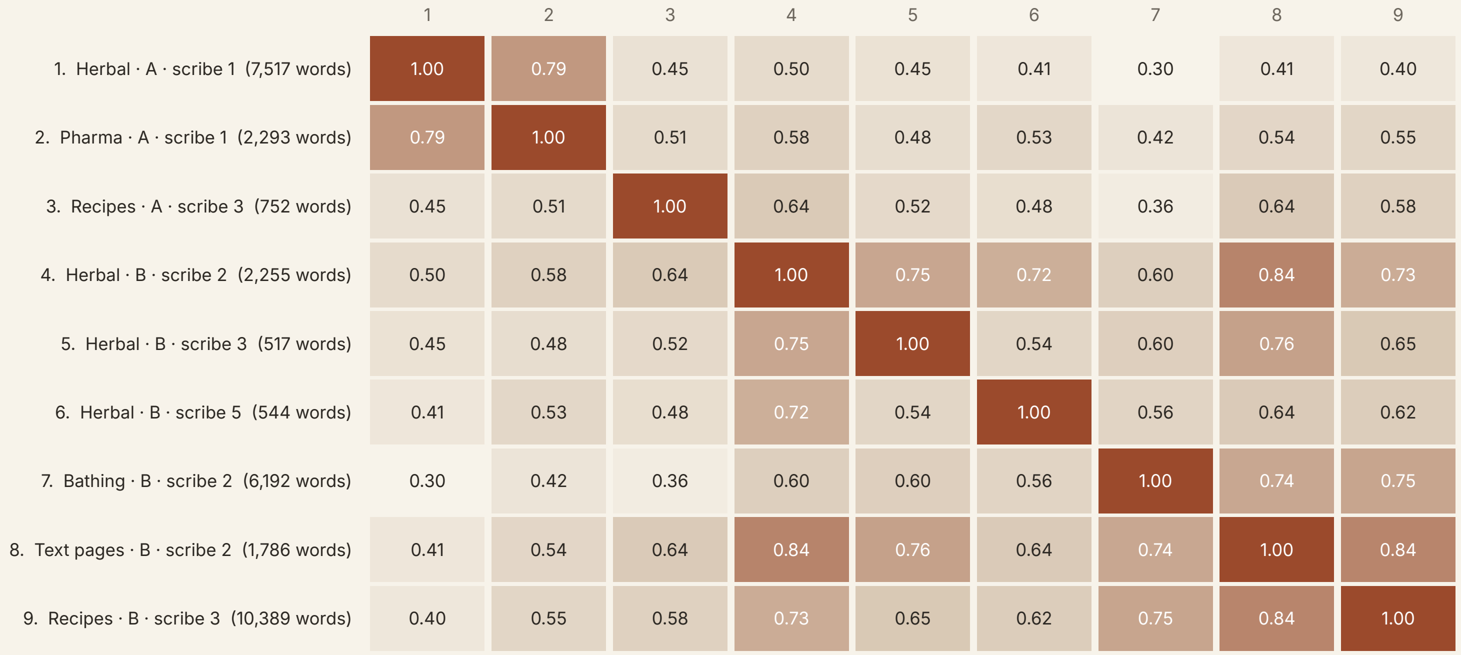
8

Two Languages, Five Hands

PRESCOTT
lys-
and Cur-
crypt-rier
was an-
a a-
crypt-lysts
an-count.
a- In

What predicts the vocabulary: subject, scribe, or language?

Similarity of word use between groups of pages defined by subject, Currier language and scribe (cosine of frequency vectors over the 400 commonest words; 1 = identical). Groups cluster first by language, then by subject; scribe matters least.



	1	2	3	4	5	6	7	8	9
1. Herbal · A · scribe 1 (7,517 words)	1.00	0.79	0.45	0.50	0.45	0.41	0.30	0.41	0.40
2. Pharma · A · scribe 1 (2,293 words)	0.79	1.00	0.51	0.58	0.48	0.53	0.42	0.54	0.55
3. Recipes · A · scribe 3 (752 words)	0.45	0.51	1.00	0.64	0.52	0.48	0.36	0.64	0.58
4. Herbal · B · scribe 2 (2,255 words)	0.50	0.58	0.64	1.00	0.75	0.72	0.60	0.84	0.73
5. Herbal · B · scribe 3 (517 words)	0.45	0.48	0.52	0.75	1.00	0.54	0.60	0.76	0.65
6. Herbal · B · scribe 5 (544 words)	0.41	0.53	0.48	0.72	0.54	1.00	0.56	0.64	0.62
7. Bathing · B · scribe 2 (6,192 words)	0.30	0.42	0.36	0.60	0.60	0.56	1.00	0.74	0.75
8. Text pages · B · scribe 2 (1,786 words)	0.41	0.54	0.64	0.84	0.76	0.64	0.74	1.00	0.84
9. Recipes · B · scribe 3 (10,389 words)	0.40	0.55	0.58	0.73	0.65	0.62	0.75	0.84	1.00

Source: ZL3b running text; groups with at least 500 words. Computation: the authors.

Figure 8.3 Vocabulary similarity between groups defined by subject, language and scribe.

the 1970s, working from the

First Study Group's transcription,

he counted letter pairs page by page and found that the pages fell into two groups so distinct that, in his words, they were “two different languages” (Currier, Two Languages, 1976). He called them A and B. He also observed that the two groups were written in two different hands, and that the hands and the languages went together. Both findings have held up for fifty years, and the second has been sharpened in 2020 Lisa Fagin Davis showed that the hands are not two but five (Davis, 2020). The question this chapter addresses is, what A and B are, and the answer, as far as the data go, is that they are not what Currier's word suggests:

Figure 8.4 Controlled contrasts.

THE DIVISION

Figure 8.1 shows the words that most separate the two groups. Three of the ten commonest words in the manuscript, **chedy**, **shedy** and **qokeedy**, occur 486, 419 and 303 times on B pages and once, twice, and never on A pages. In the other direction **chol**, **chor** and **cthy** are common in A and rare in B, and **daiin**, the commonest word in the book, is three times as frequent in A as in B. The difference is not in a few words. It runs through the spelling: the letter pair **ed** occurs in 14 per cent of all words but is almost confined to B; words ending in **-edy** and **-eedy** are B words, words ending in **-ol** and **-or** lean to A.

Because the difference is in the spelling of common words, a score can be assigned to any page. Figure 8.2 does this for every page with at least thirty words of running text, in bound order, using a crude index: the share of words with B-type features minus the share with A-type features. The herbal section is mostly A, with B pages appearing as isolated spikes in quires 4 to 7 and again in quire 17. The bathing section is uniformly and strongly B. The recipe section at the end is B, but less strongly. The pharmaceutical pages are A. The two small blocks of text pages, folios 58 and 85–86, are A and B respectively.

Two things are evident from the figure. The first is that A and B are not a continuum: pages are one or the other, and the crude score separates them

Two Languages,

nearly as cleanly as Currier's own assignments, shown as black marks beneath the bars. The second one is that B is not one thing. The bathing pages and the recipe pages are both B, but their scores differ, and so does their vocabulary: **qokain**, **qol** and **shedy** dominate the bathing section, **aiin**, **ar**, **al** and **okeey** the recipes. Currier's two languages are better thought of as a family of registers of which the herbal A and the bathing B are the extremes.

THE SCRIBES

Davis's five scribes map onto the languages but not onto the sections. Scribe 1 wrote the A pages of the herbal section and all of the pharmaceutical pages Scribe 2 wrote the B pages scattered through the herbal quires, the whole of the bathing section, and the text pages of the rosettes sheet. Scribe 3 wrote the recipe section and the B herbal pages of quire 17, and also folio 58, a page of text with marginal stars which is in language A. Scribe 4 wrote the zodiac and astronomical pages whose text is mostly labels and circles and was not assigned a language by Currier. Scribe 5 wrote a few

herbal B pages in quires 6 and 8. So, A is the work of scribes 1 and, once, 3; B is the work of scribes 2, 3 and 5 (Davis, 2020; Zandbergen, 1990–2020). This raises the obvious question: Are A and B properties of the text, or just the habits of different writers? If Scribe 2 simply preferred **chedy** where Scribe 1 preferred **chol**, the "two languages" would be two handwritings, and would tell us nothing about content.

SEPARATING THE CAUSES

The manual allows a partial test, because its pages vary in three things at once—scribe, language, and subject—and for a few combinations two of the three are held constant while the third changes. Figure 8.4 shows the results. The measure is the share of words that lie in vocabulary sig-

Two Languages,

nificantly skewed between the two groups being compared.

Hold scribe and language constant and change the subject: Scribe 1’s herbal pages against Scribe 1’s pharmaceutical pages, both in A, differ in 15 per cent of their words; Scribe 2’s herbal pages against Scribe 2’s bathing pages, both in B, differ in 26 per cent.

The subject of a page, or something that travels with it, changes the vocabulary even when the same person is writing in the same language.

Hold subject and language constant and change the scribe: Scribe 2's herbal B pages against Scribe 5's herbal B pages differ in less than 1 per cent of their words.

Two Languages,

The sample is small, 544 words on Scribe 5’s side, so the result is weak, but it is in the opposite direction from the handwriting hypothesis. Two scribes writing B on the same subject write the same words.

Hold subject constant and change both language and scribe: Scribe 1’s herbal A against

Scribe 2's herbal. B differs in 25 per cent of their words, about as much as the two subjects within one scribe.

Figure 8.3 puts all nine groups large enough to compare in one grid. The groups cluster first by language—the three A groups together, the six B groups together—and within B by subject, with the bathing and recipe pages as the two poles and the herbal B pages between them. The scribe adds little once language and subject are known.

WHAT A AND B ARE

The pattern argues against the simplest explanations on both sides. A and B are not merely two handwritings, because two scribes writing the same language on the same subject agree, and one scribe writing two subjects does not. Nor are they merely two subjects because the same subject, plants, exists in both, with the full A–B difference between its two halves. What is left is that A and B are two ways of writing the same system that were in use at the same time in the same workshop, assigned to scribes and to sections together. This is consistent with a number of readings: two dialects of a language, two spelling conventions, two settings of a cipher, two stages in the development of a method. It is also consistent with Davis's suggestions from the handwriting, that the manuscript was produced by a community of practice rather than a single author—several people who had learned the same system and used it slightly differently (Davis, 2025). Bowern and Lindemann report a related observation from the character level:

the conditional entropies of A and B differ, 2.17 against 2.01 bits, but become nearly identical, 2.23 and 2.24, if the prefix **qo-** and the suffix **-dy** are removed (Bowern and Lindemann, 2021). Much of what makes B look different from A is the heavier use of two affixes.

Montemurro and Zanette showed in 2013, using an information-theoretic measure of how words cluster, that the manuscript's vocabulary is organised by section, in the

May that the vocabulary of a real book on several subjects is: words characteristic of the herbal section are rare elsewhere, and so

Two Languages,

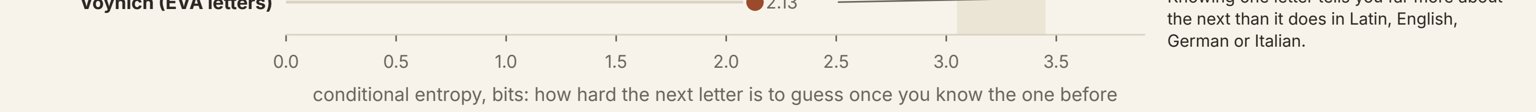
on (Montemurro and Zanette, 2013) Our contrasts confirm this with the scribe controlled. Whatever the words are, they are distributed as though they were about something.

9

The Numbers

A

Text that cannot be read can still be measured, and the measurements can be compared with those of texts that can.



Knowing one letter tells you far more about the next than it does in Latin, English, German or Italian.

Source: computed by the authors on the ZL3b transliteration and equal-length samples of the Vulgate, Pliny’s Natural History, the King James Bible, and German and Italian Bibles.

Figure 9.1 Conditional entropy against five reference texts.

This chapter collects the measures that have proved most informative, computed afresh from the ZL3b transliteration and from five reference texts of exactly the same length: a stretch of the Latin Vulgate, a stretch of Pliny’s

Natural History and passages of the Bible in English, German and Italian. The references are not meant to be candidates for the underlying language. They are there to show what ordinary text does, so that the manuscript's departures from it are visible.

Figure 9.2 Repetition and near-repetition.

ZIPF

In any large text the commonest word is about twice as frequent as the second, three times as frequent as the third, and so on: plotted on logarithmic axes, rank against frequency is a straight line of slope close to −1. This is Zipf’s law, and it is true of every natural language, of most constructed languages, and, as it happens, of a good deal of nonsense Voynichese obeys it (figure 6.3). The slope over the first thousand ranks is −1.04; the reference texts give −0.81 to −1.18. Gabriel Landini showed this in 2001 and it has been confirmed many times since (Landini, 2001). The law is a weak test: a text that failed it would be very strange, but passing it excludes little.

The shape of the curve carries more information than the slope.

The manuscript's commonest word, **daiin**, accounts for 2.2 per cent of the running text. In the English sample, *the* accounts for 9.6 per cent, in the German *und* for 7 per cent, in the Latin *et* for 7.6. Voynichese has no word that dominates in the way a definite article or a conjunction does; its top of the list is flat. This is one of several signs that whatever the words are, the commonest of them are not function words of the European kind, or that function words are hidden inside other words.

VOCABULARY

In 34,180 words of running text there are 7,466 distinct spellings, and 5,266 of them occur once. The ratio of distinct words to total words, 0.22, is higher than the English sample (0.07) or the German (0.10) and close to the Latin Vulgate (0.20), only Pliny, with his rich vocabulary of names and things, is higher (0.37). The hapax rate, 71 per cent of distinct spellings occurring once, is higher than any of the references except Pliny.

A high rate of unique spellings has two possible causes. One is a large vocabulary, an encyclopaedic text full of names. The other is spelling variation: many of the “distinct” words may be variants of one another. The second is certainly part of the story, because so many of the unique spellings differ from common words by a single letter. We return to this below.

ENTROPY

The most discussed statistic, and the one that most sharply separates the manuscript from ordinary writing, is the conditional entropy of its characters: how much the next letter is in doubt once the current one is known. William Bennett computed it in 1976 and found it far below that of any European language he tried (Bennett, 1976). Bowern and Lindemann repeated the comparison against 250 languages and found it below all of them (Bowern and Lindemann, 2021; Lindemann and Bowern, 2021). Figure 9.1 shows our own version. The single-character entropy of Voynichese, 3.88 bits, is close to that of the references (4.03 to 4.06): the symbols are used about as unevenly as letters are. The conditional entropy is

2.13 bits. The references lie between 3.11 and 3.42. Treating the benches and gallows-benches as single symbols raises the figure to 2.26, which does not close the gap. In plain terms: in the Latin sample, knowing one letter cuts the uncertainty about the next by about 0.7 bits; in Voynichese, by 1.75 bits. The next letter is far more predictable than in any natural language written plainly. Chapter 6 showed where this comes from. Most symbols have a fixed position in the word, so that once one knows the current symbol one knows roughly where in the word one is, and that restricts what can come next. A **q** is followed by **u**, an **i** by **t**, **o**, **n** or **r**; an **e** by **e**, **d**, **y** or **o**. The low entropy is the lost structure seen from a differentangle.

What the entropy rules out is any simple letter-for-letter substitution of a European language. A text in Latin with each letter replaced by one symbol has Latin's conditional entropy exactly, 3.3 bits, not 2.1. Bowern and Lindemann tried the obvious manipulations—removing vowels, abbreviating, replacing letters by syllables—and found that none of them brings a natural text down to Voynichese levels, only "systematic conflation of distinctions," such as collapsing all vowels into one symbol, comes close (Bowern and Lindemann, 2021). Conversely a sipher that replaces each plaintext letter by a *group* of symbols with internal structure, a verbose cipher, lowers the entropy naturally, because the symbols within a group are predictable from each other. That is the direction in which this statistic points, and it is the direction Greshko's

2025 cipher takes (Greshenko, 2025)

INDEX OF COINCIDENCE

The index of coincidence, the probability that two symbols chosen at random from a text are the same, is a tool of the cryptanalyst, used to distinguish simple substitution from polyalphabetic ciphers. For

Voynichese on the EVA letters it is 0.077; for the references 0.070 to 0.074. The figure depends heavily on the transliteration and is not very diagnostic here, except in one respect: it is nowhere near the 0.04 of a text enciphered with a changing alphabet. Whatever the system is, it is consistent: the same thing is written the same way throughout.

REPETITION

The feature of the text that most struck its early readers, and that remains the hardest to explain, is how often it repeats itself. Figure 9.2 quantifies it. In the running text, 0.8 per cent of words are identical to the word immediately before them: **daiin daiin**, **chol chol**, **qokeedy qokeedy**. In the reference text the figure is below 0.1 per cent. A further 3.3 per cent of words differ from the previous word by a single letter, against 0.3 to 1.2 per cent in the references. Within a window of five words, 16 per cent of Voynichese words have a near-twin, which is more than the Latin (4–7 per cent) and the German and Italian (10–14) and about the same as the English sample, whose repetitiveness is the King James Bible's. The baseline matters.

If the same 34,180 Voynichese words are shuffled into random order, the adjacent-repeat rate falls to 0.3 per cent and the near-repeat rate to 1.7 per cent.

So the repetition is not merely a consequence of a small vocabulary; the text is locally far more repetitive than its own words.

in random order. Torsten Timm and Andreas Schinner built a theory of the manuscript on this observation: they propose that the text was generated by copying and slightly modifying words already on the page, a procedure they call auto-copying, which reproduces the near-repeat rate, the slot structure and the line effects without any underlying message (Timm and Schinner, 2020). Chapter 16 weighs that proposal. The observation itself is secure.

LONG-RANGE ORDER

The repetition is local. There is also structure at long range. Montemurro and Zanette computed, for each word, how much information its distribution across the text carries—whether it clusters in some parts and is absent from others, as the keywords of a book on several subjects do—and found that the manuscript's words cluster in this way to about the degree that words in real books do, and that the clustering follows the pictorial sections (Montemuro and Zanette, 2013). Our controlled contrasts in chapter 8 are the same finding with the scribe held constant. Amancio and colleagues, using network measures of word co-occurrence, likewise placed the manuscript among meaningful texts rather than among the shuffled versions they used as controls (Amancio et al., 2013).

THE SUM

Set side by side, the statistics describe a text that is more regular than language at the level of the symbol and about as regular as language at the level of the word and the section, with a layer of local repetition in between that language does not have. Any explanation has to produce all three. Simple substitution of a European language fails at the first. Random generation fails at the third. The candidates that survive are the ones in which the symbols are not letters: a verbose or homophonic cipher, a system of abbreviation, a syllabary or similar script for an unknown tongue, or a very well-designed generator. Those are the subject of Part IV.

Part
III

The Pictures

10

Plants

ABOUT half the book is plants: 129 pages in the ZL38 count, nearly all in the first seven quires, with a scattering later (Zandbergen, 1996–2026). The format is as regular as a field guide. One plant to a page, drawn large, usually with root, stem, leaves and flower all shown; a block of text above the plant or beside it, sometimes continuing below; occasionally two plants, and occasionally a second block of text. The drawing was made first, in ink, and the text was fitted around it, which is the order in which illustrated herbals were normally made. The text on a herbal page averages 84 words. Some pages have a few lines, one has 163 words, and a handful have no text at all.

THE GENRE

The illustrated herbal is one of the old-est kinds of book there is, and by the fifteenth century it had a long tradition behind it, descending through Latin translations of Dioscorides and the pseudo-Apuleius herbal to the *Tractatus de herbis* family of the fourteenth and fifteenth centuries, in which each plant has its picture and a paragraph of properties and uses (Clements, 2016). Anyone who has looked at a *Tractatus de herbis* manuscript recognises the Voynich herbal pages instantly as the same kind of thing: the same one-plant-per-page layout, the same emphasis on the root, the same habit of drawing the plant as a diagram of its parts rather than as it grows. Historians of botany who have examined it place it in that tradition while noting that it copies no known exemplar (Clemens, 2016; Zandbergen, 1996–2026).

That last point is the difficulty. In most herbals the pictures were copied from earlier herbals, and a plant can be identified by tracing its picture back through the copies to one that was labeled. The Voynich plants have mostly resisted this. A few look like real plants—there is a plausible water lily, several that could be members of the daisy family, something that looks like a viola—and one, on folio

35v) has a close match in a Paris manuscript of the *Tractatus de herbis* (Bibliothèque nationale, lat. 6823, folio 60r), noticed by several researchers (Zandbergen, 1996–2026). Many others combine the root of one plant with the leaves of another, and the flower of a third, or have roots drawn as if they were animals or faces. Theodore Petersen and Ethel Voynich, in the 1930s, each compiled lists of tentative identifications; the lists agree with each other more than with later botanists, and more with recent amateur identifications in some cases, but no list has won general acceptance (Zandbergen, 1996–2026).

Two cautions apply to all such attempts. The first is the paint. As chapter 4 showed, the colours were at least partly applied after the pages had been shuffled, by someone who did not keep inside the lines, so a blue flower may be blue because the painter had blue. The second is that the identification of a plant is only useful if it leads somewhere: if the label or the text beside it can be shown to contain the plant's name in some language. No identification has yet done that.

THE SUNFLOWER

The best-known identification is the worst. In 1944 the botanist

Hugh O'Neill published a note identifying the plant on folio 93r as a sunflower and another as a capsicum pepper, both American plants, and concluded that the manuscript must postdate 1493 (O'Neill, 1944). The claim was widely repeated for sixty years. The radiocarbon date of 1404–1438 ends it: whatever, the plant on folio 93r is, it was drawn before Columbus sailed. The episode is instructive, because O'Neill was a competent botanist and his sunflower looks, to a modern eye, like a sunflower. The drawings are suggestive enough to support almost any identification, and that is exactly why identifications without independent support do not count.

LABELS AND TEXT

Unlike the pharmaceutical pages, the herbal pages carry almost no labels: five words in the whole section, on three pages. Whatever names the plants have are in the paragraphs, not beside the drawings. The paragraphs themselves are the longest and most regular text in the book: 247 of the 736 paragraphs are here, and the paragraph-opening gallows rule of chapter 7 is at

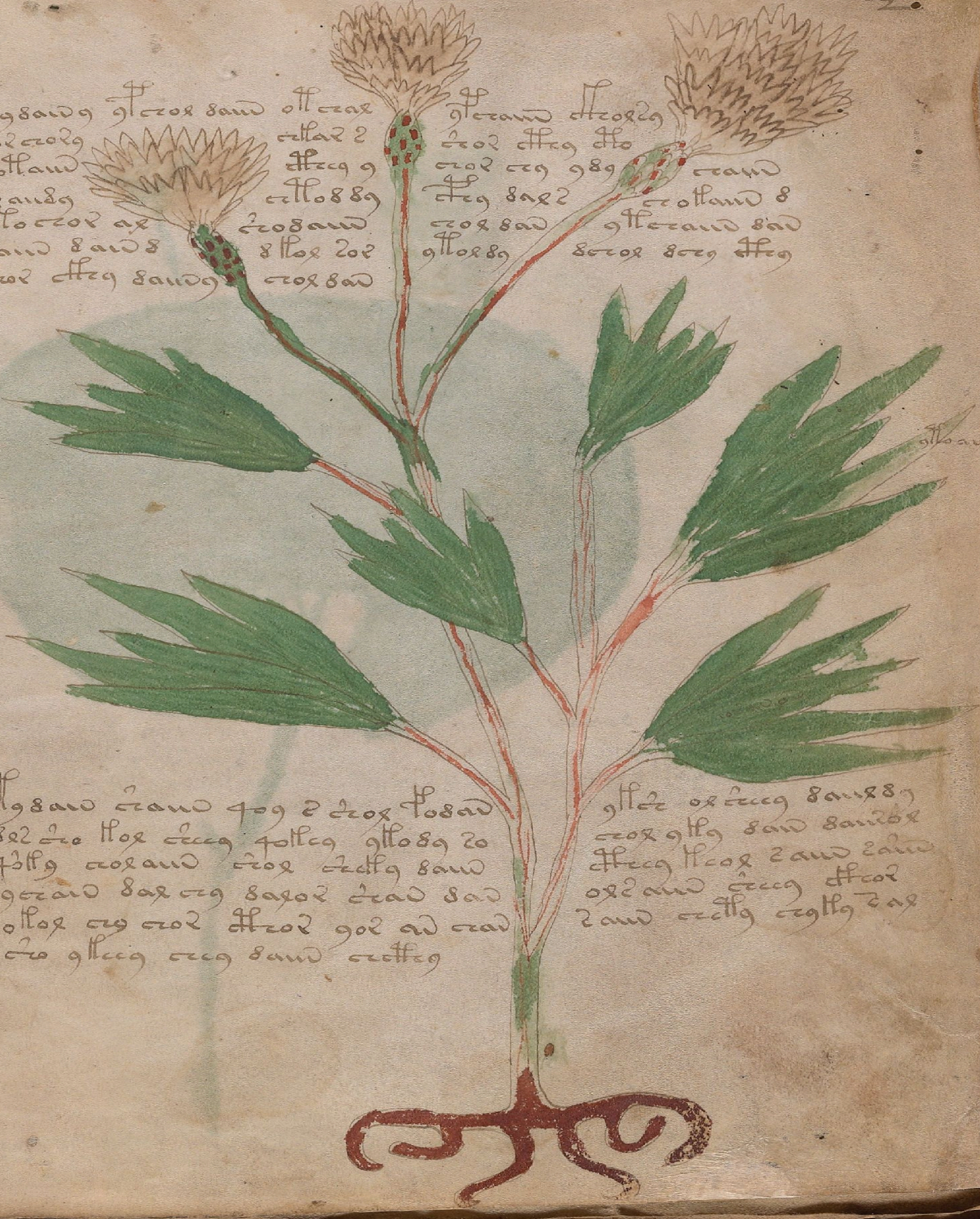
its strongest. In Currier's A pages the vocabulary is dominated by **daiin**, **chol**, **chor**, **shol**, **cthy**; the thirty-two B pages mixed among them use **chedy**, **shedy**, **qokedy**, as the bathing section does.

If the text describes the plants, the descriptions should have something in common with one another, and with no other section,

Montemurro and Zanette's analysis found exactly this: the herbal section has its own characteristic vocabulary, and the words that character it are among the strongest keywords in the manuscript (Montemuro and Zanetta, 2013). Our contrast in chapter 8 confirmed that Scribe 1's herbal pages differ in vocabulary from his pharmaceutical pages. That is what a herbal text would do. It is also what a well-made imitation of a herbal text would do, if its makers varied the vocabulary section by section, and we do not know which we are looking at.

ORDERING

A real herbal is usually ordered: alphabetically by plant name, by the part of the body treated, by the plant's qualities, or following an older model's order. Order would give a way in; if the herbal were alphabetical, the first words or labels of successive pages would begin with the same letters in sequence. No such sequence is visible in the first words of the herbal paragraphs, which begin with the paragraph-initial gallows words, nor in the pattern of vocabulary across pages. Given that the sheets are out of order, this is less informative than it would otherwise be, and is one more reason to hope that the original sequence can be recovered.



f. 2r. A typical herbal page: a single plant drawn with root, stem, leaves and flowers, and a block of text above. Beinecke Rare Book & Manuscript Library, Yale University

f. 25v: the plant with its closest known match in another herbal, Paris BnF lat. 6823, fol. 60r.

f. 41v: one of the thirty-two herbal pages in Currier's language B, written by Scribe 1. Beinecke Rare Book & Manuscript Library, Yale University.

11

Stars and the Zodiac

QUIRES 9 to 12 are the strangest part of the book to handle. These of the four are single large sheets folded two, three or four times, so that opening them turns the small book into a broadsheet, one quire is a single ordinary bifolium. All of it is by one scribe, Davis’s Scribe 4, who wrote nothing else, and almost none of it is paragraph text: of the 2,168 words these quires, 276 are in paragraphs, the rest in labels and in rings of text running around circles. Currier did not assign these pages to either of his languages.

THE ASTRONOMICAL SHEETS

Quire 9 is one sheet with six panels, folios 67 and 68. On them are circular diagrams with a sun or moon face at the centre, rings of text, and scattered stars, each star with a label. One panel, folio 68r3, shows a cluster of seven stars with a label, which has been read since the 1920s as the Pleiades, the one asterism in the sky that a medieval reader would draw as seven dots in a bunch, next to it, a crescent moon. The identification is plausible and unprovable. Another parcel, folio 68v3, has a circle divided into one half and two quarters, the T-O pattern that medieval cartography used to show the three continents (Zandbergen, 1996–2026). Quire 10, folios 69 and 70, holds further circular diagrams of the same family, with radiating text and stars.

Nothing in these diagrams has been matched to a known astronomical table or figure. There are no recognisable constellation outlines, no coordinates, no planets. The star labels, 86 words, are nearly all unique, which is what star names would be; but **dchol**, the one that recurs, is also an ordinary herbal word, and more than half of the star labels occur in the running text.

THE ZODIAC

The zodiac occupies the rest of quire 10 and quires 11, and

12: twelve panels, folios 70v to 73v. Each is a wheel. At the centre is an emblem: two fishes, a ram, a bull, a pair of standing figures, a crayfish-like creature, a lion, a woman, a balance, a creature usually taken for a scorpion, an archer with a crossbow. Around the emblem run two or three rings of small female figures, most of them in tubs, each holding a star on a string, and beside each figure a label. The rings are interrupted by bands of circular text.

The sequence has three peculiarities, all noted by Zandbergen (Zandbergen, 1996–2026). It begins with Pisces rather than Aries, which is almost unknown in medieval zodiacs. Aries and Taurus each appear twice, on two panels with fifteen figures each, one with the animal drawn dark and one in light, so that the two halves together make thirty; every other sign has its thirty figures on one panel (Pisces has twenty-nine). And Capricorn and Aquarius are missing, together with folio 74, which is exactly where they would have been.

Thirty figures to a sign is the number of degrees in a sign, and the obvious reading, which Zandbergen and most others accept, is that each figure stands for one degree (Zamberger, 1996–2016). Medieval astrology had a literature of the degrees, the *paranatellonta*, in which each degree of each sign has its own image and meaning, descending from Hellenistic sources through Arabic intermediaries to Latin manuscripts of the kind that circulated in the fifteenth century. The Voynich wheels do not resemble any surviving degree-image series closely, and the figures are nearly interchangeable; what varies from figure to figure is the label.

THE MONTH NAMES

Here, and only here, the manuscript contains writing that can be read. In the centre of each zodiac panel, beside the emblem, someone has written the name of a month in Latin letters; March beside the fishes, April beside both rams, May beside both bulls, and so on to December beside the archer. The hand is not that of the text; the ink and the letter forms are later, and the forms are those of a Romance vernacular. The precise dialect has been argued over since the 1990s—French, Occitan, and northern Italian forms have all been proposed, and some of the spellings are unusual—and Zandbergen describes them simply as “a version of French” (Zandbergen, 1996–2026). The association of Pisces with March, Aries with April and so on is the ordinary medieval one, in which a sign is named for the month in which the sun enters it.

What the month names tell us is limited but real. Someone who could read the Latin alphabet and spoke a Romance language had the book, understood the emblems well enough to assign months to them, and did not add anything else: no plain names in the herbal, no title, no glossa-

on a single Voynichese word. They are the earliest readers we know of, and they could not read it either.

THE LABELS

The 345 zodias labels are the largest set of labels in the book and the most tempting, because they stand beside figures that almost certainly represent something specific: a degree, or a star, or a day. If they were names, in whatever system the text uses, a decipherer would have 281 distinct names with known referents.

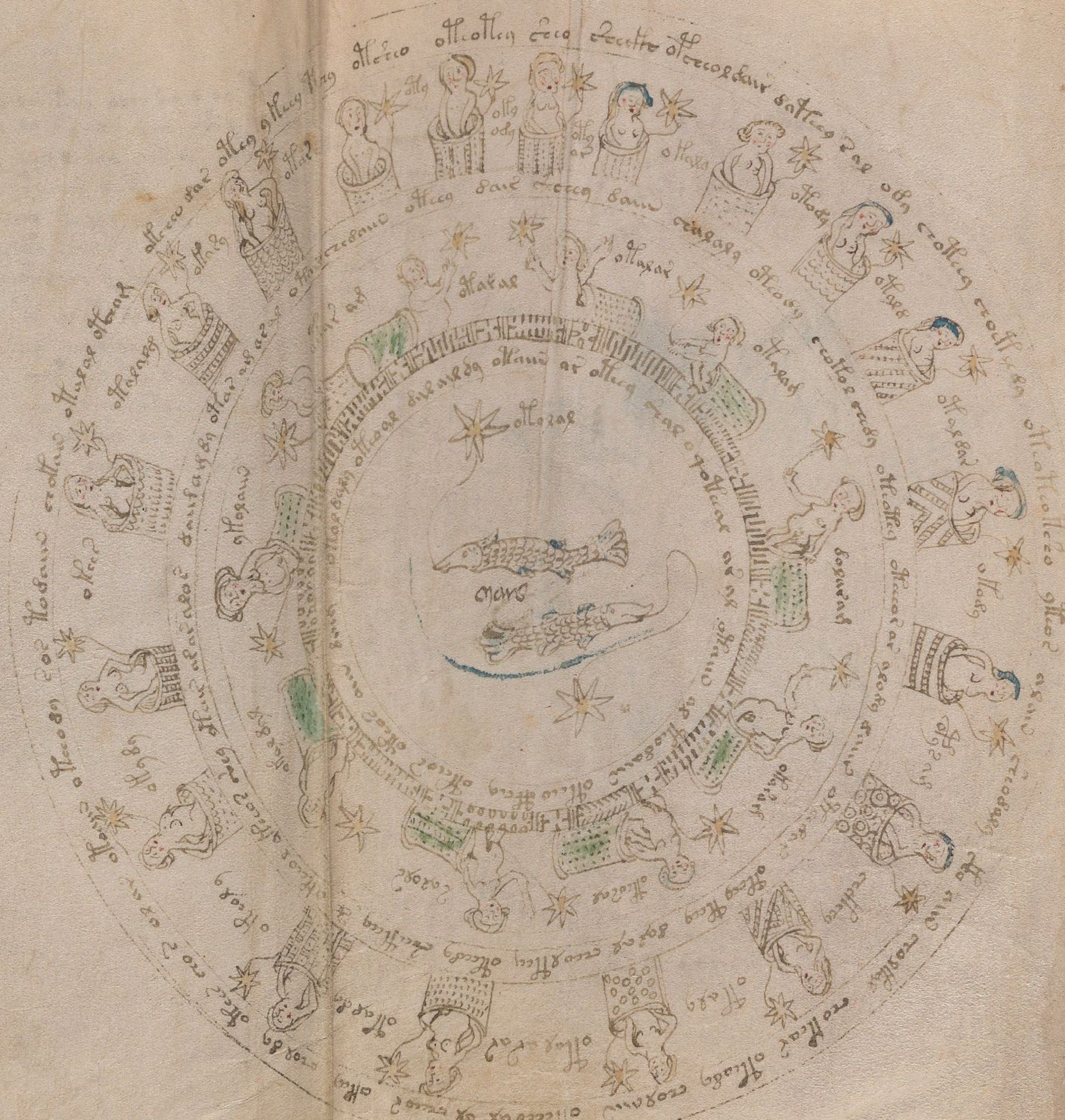
The distribution argues against the simplest version of this. Thirty-seven of the 281 distinct labels occur on more than one zodiac panel: **oky**, **al**, **otaly**, **otal** and **okeey** each appear on four different signs. Names of degrees would not repeat across signs; names of stars would not either. And the labels are made of the ordinary vocabulary of the running text—over half occur there as words rather than looking like a separate onomastic stock. Whatever they are, they are not thirty unique proper names per sign. They might be attributes of the figures, or abbreviated phrases, or something else, the evidence does not say.

COMPARANDA

The emblems themselves have been compared with zodiac cycles in hundreds of other manuscripts. The closest resemblances, in the crossbow of Sagittarius, the odd Scorpio, and the dress and headgear of the human figures, are with manuscripts from southern Germany of about 1400 to 1500 (Zandberger, 1996–2016). This agrees with the radiocarbon date and with the one other datable feature of the drawings. It does not identify a source; no known zodiac has the Voynich arrangement of tubs and rings, and a full cycle begin-

ning
with
Pisces
has
no

known
par-
al-
lel.


mars

f. 70v2, Pisces: the first sign of the cycle, with twenty-nine figures in two rings and the month name written beside the fishes in a later hand. Beinecke Rare Book & Manuscript

f. 70v1, the first of two Aries panels, with fifteen figures. Beinecke Rare Book & Manuscript Library, Yale University.

f. 73v, Sagittarius with a crossbow, the feature most often cited for a Southern German origin. Beinecke Rare Book & Manuscript Library, Yale University.

Stars and the

f. 68r, the astronomical foldout: rings of stars with labels, a sun and a moon, and at the right the cluster of seven stars often taken for

Stars and the

the Pleiades. Beinecke Rare Book & Manuscript Library, Yale University.

12

The Bathers

QUIRE 13 is the most coherent gathering in the book. Twenty pages, folios 75 to 84; one scribe, Davis's Scribe 2; one language, the most extreme form of Currier's B; one subject. The subject is women in water. Page after page shows small, plump, mostly naked female figures standing in green pools, sitting in tubs, lying in channels, holding on to pipes. The pools are connected by tubes and conduits that run between paragraphs of text, so that the whole quire reads as a plunging diagram with people in it. There are more than two hundred figures here, and over five hundred

in the book as a whole (Zandbergen, 1996–2026). The section has been called biological, balneological, anatomical and gynaecological by different writers, we call it the bathing section, because that is what the pictures show, and leave what it is about open.

WHAT IS DRAWN

The figures are of a single type, the same type that holds the stars in the zodiac wheels: round-bodied, round-faced, with loose hair or a head-dress, and a stylised posture. A few hold objects, among them rings and what may be flowers. The water is drawn in green and blue, sometimes with wavy lines

for flow. The vessels range from plain basins to elaborate structures with crenellated rims and arches, and the pipes include spouts, funnels and what look like valves or taps. On folio 77r figures stand beneath a band of coloured arcs across the top of the page and down its left margin each attached to a pipe;

on folio 78r two pools, one above the other, are fed from a structure at the head of the page; on folio 84r a pool at the top is roofed with a row of figures and tubs, like a gallery.

Whether the pipes are pipes is a fair question. The drawings have been read as hydraulic engineering, as the vessels of the body with the figures as humours or spirits, as alchemical apparatus, and as the plumbing of a bathhouse.

Zandbergen notes the resemblance of some of the objects to those in illustrated copies of the *De balneis Puteolanis*, a thirteenth-century poem on the healing baths near Naples whose manuscripts show bathers in pools and the equipment of the bath (Zandbergen, 1996–2026). Bathing as medicine—balneotherapy—was an established branch of medieval practice, and a treatise on the baths with pictures of bathers would not be out of place beside a herbal and a pharmacy. Beyond that

resemblance of genre, no specific source has been found.

THE TEXT

The bathing section contains 6,192 words of running text in 86 paragraphs, more text per page than any other illustrated section, and it is where the B vocabulary is at its purest. **qokain**, **qol**, **qokeedy**, **shedy**, **qokal**, **shey**, **qokeey** and **qoky** are all far commoner here than anywhere else; **daiin**, **chol** and **chor** nearly vanish. The prefix **qo-** is at its most frequent. Chapter 8 showed that these differences hold even against the same scribe's herbal pages: whatever Scribe 2 was writing here, he wrote it differently from what he wrote beside the plants.

The

64 labels beside figures and 55 beside tubes and vessels follow the general rule for labels: rarely q-initial, mostly o-initial, and drawn from the

ordinary vocabulary, of which four fifths of the figure labels are members. **otedy**, **otol**, **dal** and **daldy** each label more than one figure.

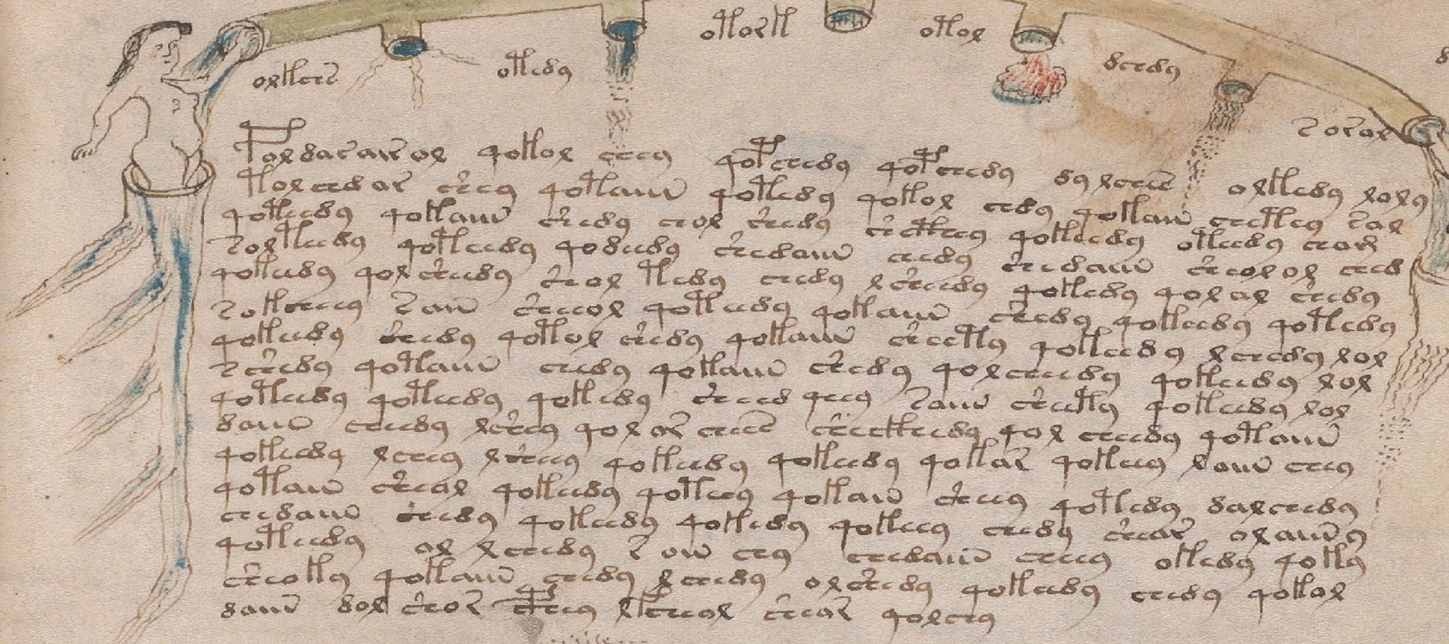
FOLIO 78 AND FOLIO 81 The section contains the single most useful piece of evidence about the book's order, described in chapter 4: a stream of water drawn from a vessel at the right edge of folio 78v continues from the left edge of folio 81r. The two pages belong to one sheet and were drawn as one opening. Another sheet, folios 79 and 80, is now bound inside them, so that the picture is broken. The sheet 78/81 was originally the innermost of its gathering (Davis, 2025; Zandbergen, 1996–2026). This is also the clearest case where the text might one day help if

the paragraphs on 78v and 81r are continuous, a reconstruction that joins them should show it, and Layfield and Davis's similarity measure is one way to look.

WHAT CAN BE SAID

Three things. The section is a planned unit: one sheet-maker, one scribe, one vocabulary, one iconography, drawn and written together. It is the part of the book whose text behaves least like the herbal text, which is some evidence that the text follows the pictures, that is, that it is about something. And it belongs to a recognisable medieval genre, the illustrated treatise on bathing and the body, without being a copy of any known example. The rest—what the women are doing, what the water is, why the pipes—is not something the manuscript tells us,

and
we
do
not
tell
the
reader.



f. 77r. Figures along the top and left margins, each with a pipe; paragraphs of language-B text between. Beinecke Rare Book & Manuscript Library, Yale University.

f. 78r. Two pools, fed from the structure: a pool topped by a gallery. Beinecke Rare Book & Manuscript Library, Yale University.

f. 84r. The last page of the section: at the top of the page of figure. Beinecke Rare Book & Manuscript Library, Yale University.

13

The Rosettes

QUIRE 14 is a single sheet of parchment about 45 centimetres square, folded in six, and it is the only part of the manuscript that is famous as a picture. Unfolded, one side shows nine large circles arranged three by three, each filled with a different pattern—rays, scales, stars, tubes, a ring of tiny figures—and joined to its neighbours by causeways, pipes and scrolls. The other side carries smaller circular diagrams and three panels of plain text.

The text is Currier B by Davis's Scribe 2, the scribe of the bathing section, which is one reason to think the sheet was bound near where it belongs (Davis, 2020).

WHAT IS DRAWN

Each of the nine circles is different. The central one has a ring of small towers or pavilions and a ring of small figure around a patterned core. Two circles, at opposite edges, have a starburst of blue rays. One is filled with a field of stars. One has, at its edge, a group of buildings with walls and towers whose battlements end in the forked merlons that are characteristic of northern Italian and

Alpine fortification in the late Middle Ages — the detail most often cited in arguments for an Italian or South Tyrolean origin (Zanbiergen, 1996–2026). In one corner of the sheet, outside the circles, is a small circle divided into a half and two quarters, the T-O schema of the meridian world map, which also occurs on folio 68v (Zandbergen, 1996–2026). Around and between the circles runs text: labels, rings of writing along the rims, and words along the connecting paths.

The sheet has been read as a map, a cosmological diagram, a plan of a city or a set of cities, an itinerary, and a scheme of the elements. Most readings notice the same things: that the circles are connected in a way that implies movement or relation between them; that the central circle is the largest and most elaborate; that the cascade makes one circle look like a place that the stars and rays make others look like sky. None of these readings has been shown to be more than a resemblance.

ORIENTATION

Even which way up; the sheet goes is uncertain. The folds were resewn at least once, the sheet was foliated folded, and the circles are drawn to be looked at from more than one direction, with text that runs around rims in both senses. We reproduce the sheet in the orientation of Yale's photograph, which is one of the possibilities, and describe positions on it without assuming that orientation is original.

The rosettes sheet holds 1,678 words in the ZL3b count, of which 655 are in paragraphs, on three panels of the reverse, and the rest in rings and labels. The paragraph text is B of the bathing-section kind. Of all the parts of the book, this is the one where the words most obviously stand in relation to specific drawn thing—a word beside a tower, a ring of words around a circle—and the one where that relation is least understood. It is, with the zodiac labels, the place and decipherment would be tested first.

f. 85v–f. 86r, the rosettes sheet unfolded, in the orientation of Yale’s photograph. Nine circles joined by causeways; a walled town with forked battlements at the lower right; a small T-O circle at the lower left. Beinecke Rare Book & Manuscript Library, Yale University.

14

Jars and Roots

THE pharmaceutical pages are the part of the book that looks most like a working document. There are sixteen of them in the ZL30 classification, on fold-out leaves in quires 15 and 19, folios 88, 89, 99, 100, 101 and 102. Each shows rows of small drawings of plant parts—shoots mostly, with some leaves and the occasional whole plant—laid out across the page like specimens on a bench, with a short label beside each. At the left end of each row stands a drawn container: a cylindrical vessel with a lid, a tall jar with a foot, a jar with bands and ornament. Between the rows run short paragraphs of own text. The whole section is in Currier's language A, by Scribe 1, the scribe of the herbal As pages (Davis, 2020).

THE LAYOUT

The pages are organised as a table. Reading across, a row is a container followed by a series of plant parts; reading down, the containers form a column at the margin. The natural inference, made by everyone since Newbold, is that each row lists the ingredients of some things kept in the container, or the preparations made from the plants in that row, and that the paragraphs between rows describe them. The drawings of plant parts are small and schematic but not careless: roots are drawn with their branching, leaves with their lobes, and some can be matched by eye to plants drawn large in the herbal section though no systematic concordance has been established.

The containers are the most puzzling objects in the book. Some have been compared to apothecary jars, some to reliquaries, some to the vessels of a distiller. They do not closely resemble surviving pharmacy jars of the period, which were plain earthenware; they are more ornate, with tiers and knobs, and one or two look like pieces of metalwork. They might be drawn from life, from imagination, or from a convention we do not know.

THE LABELS

This is the section where labels matter most, because there are many and they are attached to small, distinct objects. The ZL transliteration records 202 words labelling plant parts and 36 labelling containers. The plant-part labels behave like other labels: only one in two hundred begins with **q**, half begin with **o**, and half of the distinct labels also occur as words in the running text. **okol**, **otoldy**, **otoky**, **otaly**, **okaiin** and **oky** each label more than one plant part and **otaly** and **oky** also label figures in the zodiac. The container labels are different. There are 36 of them, every one unique, with an average length of seven letters against the five of ordinary words, and only three of the 36 occur anywhere in the running text. Whatever the containers are, their labels are made of the same letters in the same slots, but they are longer and rarer than anything else in the book: **otorchety**, **orfyskydal**, **daramdal**. If any class of word in the manuscript is a set of names, this is the best candidate.

WHAT CAN BE READ OFF THE PAGE

Without reading, a word, the section

tells us some things. It is a reference list, not a narrative: its unit is the row. It relates plants to containers, and so, presumably to preparations, in the manner of a pharmacopoeia or a recipe collection. Its plants are, by the drawings, mostly roots, which is where the active substances of most medieval simples were held to reside. And it was written by the same scribe as the herbal, in the same language, which suggests that herbal and pharmacy were conceived together as they are in the medical compendia of the period, where a *Tractatus de herbis* is followed by an *Antidotarium*. Zandbergen has observed that the two bifolia in quire 19 and the one in quire 15 belong together, and that on the style of the containers the quire 19 leaves should come first (Zanberger, 1996–2016).

The section is therefore one more case of a planned sequence broken by the binder.

f. 88v–f. 89r: rows of plant parts, each with a label, and a column of containers at the left of each page. Beinecke Rare Book & Manuscript Library, Yale

University.

f. 99r. Beinecke Rare Book & Manuscript Library, Yale University.

f. 190r. The leaves at quire 19, which on the style of the containers precede those of quire 15. Beinecke Rare Book & Manuscript Library, Yale University.

15

Starred Paragraphs

THE last quire is the plainest and, by word count, the largest: Twenty-three pages, folios 103 to 116, carry nothing but text: paragraph of six to twelve lines, each paragraph marked in the left margin by a small star, most of the stars with a tail, some painted red, some with a coloured cen-

tras (Zandbergen, 1996–2026). There are 292 small paragraphs in the ZL3b count, counting all folio 58 in quire 8, which has the same format and the same stars and is the only other page of its kind. The quire holds 11,602 words, a third of the running text of the whole book, and every page is by Scribe 3 in

Currier's B.

A LIST OF SOMETHING

The format is that of a list of entries, and the stars have been read as bullet points, as the mark of a recipe, and as literal stars: one proposal going back to the 1920s is that the section is a calendar or almanac with one entry per day or per star, and the association of the zodiac figures with stars gives the idea some colour. The number of paragraphs does not help: 292

is not 365, nor a multiple of 30, nor the number of stars in any catalogue, and two leaves are missing from the centre of the quire, so the original count is unknown.

What the text does say is that these entries are shorter and more uniform than the paragraphs of the herbal section, and that they are in a B that is distinctly its own. Against the bathing section, which is also B and also a single scribe, the recipe pages favor **aiin**, **ar**, **al**, **okeey**, **otaiin**, **chol** and **lkeey**, and use **qokain**, **qol** and **shedy** much less. Over a quarter of all words in the two sections lie in the vocabulary significally sketched between them. Since the scribe changes as well as the subject, the two causes cannot be separated here as they could for the herbal; but the direction is the same as everywhere else in the book. Different pictures, or no pictures, go with different words.

THE LAST PAGE

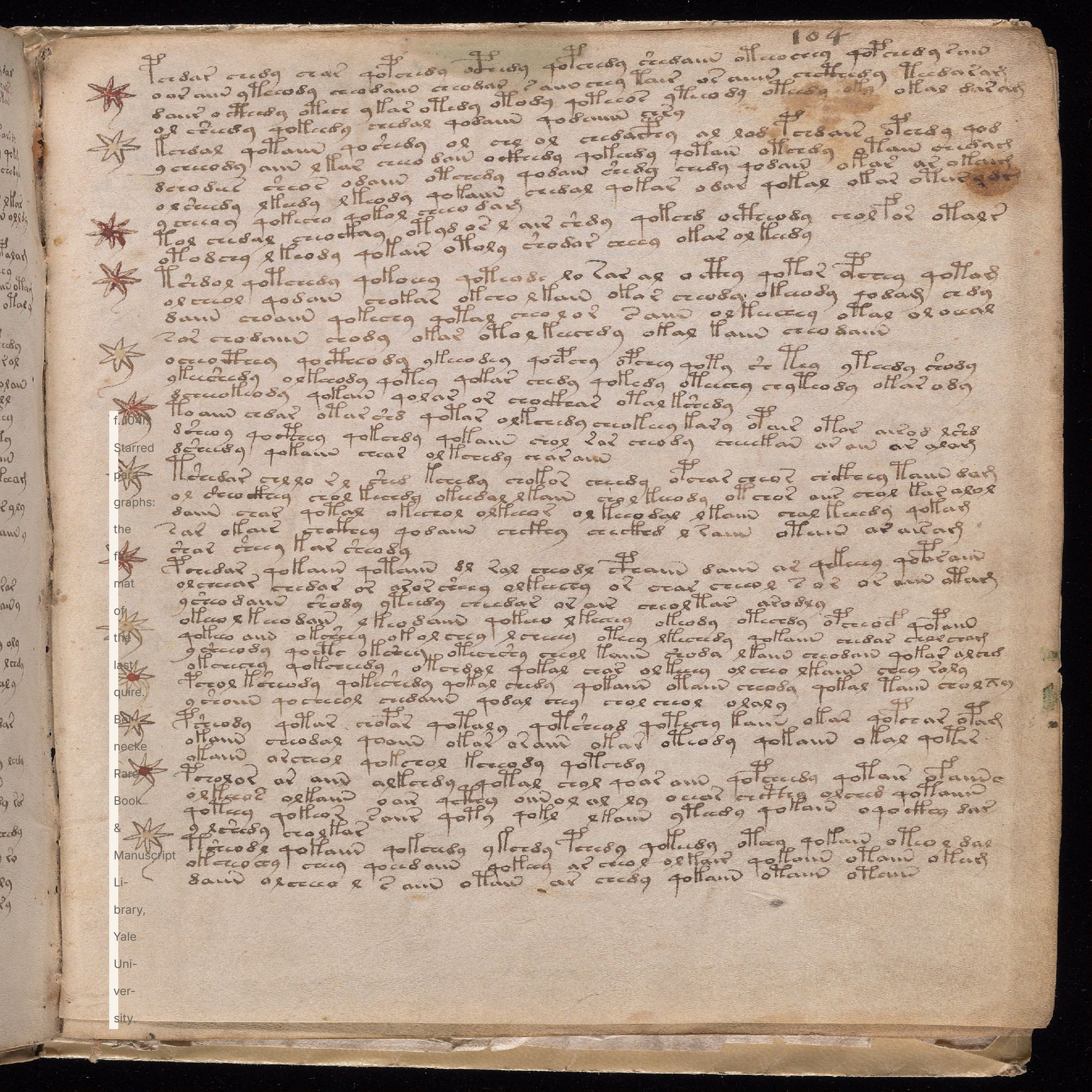
Folio 116v, the back of the last leaf, is not part of the sequence. It carries a few lines in a mixture of scripts: words in the Latin alphabet, in a hand that is not that of the main text and may be sixteenth-century, interspersed with words in Voynichese, with small drawings in the margin of an animal and a figure. The Latin alphabet words have been read in many ways, the readings agree that some of them look German (the sequence usually read as *so him gaf mich* has been taken as a German phrase) and that others, like the much-quoted *michiton oladabas*, look like nothing in particular. The lines have been interpreted as a charm, a recipe, a pen-trial and a later owner's attempt to write in the script (Zandbergen, 1996–2026). They are the only place in the book where the Latin alphabet and the Voynich script share a line, and they have not helped: whoever wrote them was not providing a key.

Two other marginal notes in Latin letters exist, on folios 17r and 66r, both short and both worn, and on folio 49v a column of letters or numerals runs down the margin beside the lines of text, in a hand that may be the same as the month names (Zandbergen, 1996–2016). None has been read with confidence.

WHAT THE SECTION OFFERS

For a decipherer the recipe section is the richest material in the book: one large body of text in a uniform format by one hand, with no pictures to complicate the layout, and with a visible unit, the paragraph, that presumably corresponds to a unit of content. If the entries are recipes, each should name ingredients that also appear in

the herbal and pharmaceutical sections, and a decipherment that got the herbal right would be checkable here. If they are almanac entries, the structure should reappear with the calendar. Either way the section is where any claimed reading of the manuscript would have to prove itself at scale.



f. 104r. Starred paragraphs: the format of the last quire. Beinecke Rare Book & Manuscript Library, Yale University.

f. 116v, the last page: lines mixing Latin letters, like those of Voynichese, with marginal drawings. Beinecke Rare Book & Manuscript Library, Yale University.

Part

IV

Explanations

16

What It Could Be

EVERY explanation of the manuscript belongs to one of four families. The text is a natural language written in an invented script. It is a natural language disguised by a cipher. It is a constructed language or artificial system of some kind. Or it is meaningless, produced to look like writing. Within each family the proposals differ in detail, and a few straddle the lines, but the four are enough to organise the evidence, and this chapter takes them in turn, holding each against what Parts I and II established. The aim is not to pick a winner. It is to see which survive.

NATURAL LANGUAGE, UNKNOWN SCRIPT

The idea is simple somebody wrote a real language, perhaps their own, in an alphabet of their own devising, and the text is as readable as any language in any script once the script is understood. This is how Linear B and the Maya glyphs were eventually read. It is the explanation most readers assumed by default, and it has been given many specific forms: Latin, Hebrew, a Romance dialect, a Turkic language, Nahuatl, a lost language of the Alps. Three findings bear on it. First, the character-level entropy. A language written out plainly in an alphabet, with one symbol per sound, has the conditional entropy of that language, which for every European language measured lies between about 3 and 3.5 bits. Voynichese has 2.1 to 2.3. No known language written plainly is that predictable; Bowern and Lindemann's survey of 250 found none (Bowern and Lindemann, 2021). So the script, if it is a script, is not a plain alphabet. It could be a syllabary, which writes consonant-plus-vowel as one symbol and so has fewer choices at each step; or an abjad, that drops vowels, although abjads normally raise the entropy, not lower it; or an alphabet with positional variants in which the same sound is written differently at the start, middle, and end of a word, as Arabic letters are. Positional variants would produce exactly the slot structure of chapter 6 (Bowern and Lindemann, 2021). Second, the word lengths. Natural languages have many short function words; Voynichese has few one- and two-letter words and a tight peak at five. A syllabary or a positional alphabet could produce this if most syllables or words were written with the same number of marks.

Third, the repetition. **daiin daiin daiin** and its cousins are hard to find in natural prose. Some languages reduplicate; some texts, like litanies, repeat. The rate here, ten to twenty times that of the reference texts, would need a genre as well as a

language. The family survives, in its syllabary and positional-variant forms, with the burden of explaining why, in a century, no one has found the language. Our own substitution experiments in chapter 17 showed that no simple mapping onto Latin, Italian, German or English works, with or with-

are vowels, but they do not test the languages no one has tried, and they cannot.

CIPHER

The second idea is that the text is a known language, almost certainly Latin or a vernacular of the region transformed by a procedure meant to hide it. This was the first explanation anyone offered—Baresch, Kircher’s correspondents, and the Yale catalogue all call the book a cipher—and the professional cryptanalysts of the twentieth century, Friedman, Tiltman, Currier, all approached it as one and all failed.

Ciphers of the early fifteenth century are well documented. Simple substitution was the rule; nomenclators, which replaced common words and names by symbols, were coming in; transposition was rare; polyalphabetic ciphers were still half a century away. The entropy excludes simple substitution outright. It also makes polyalphabetic substitution unlikely, since such ciphers flatten the statistics rather than sharpening them, and the index of coincidence knows nothing of the kind. A nomenclator, alone would leave the ordinary text readable.

What the statistics are compatible with is a verbose cipher: a scheme in which each letter of the plaintext is replaced not by one symbol but by a short group drawn from a table, so that several Voynichese glyphs correspond to one plaintext letter. Such a scheme lowers the entropy, because the glyphs within a group predict one another; it produces words of nearly uniform length, because the groups are of nearly uniform length; and it produces the slot structure, because each group has a shape. Michael Greshko's Naibbe cipher of 2025 is a worked example using tables that could be chosen with a deck of playing cards, it turns Latin or Italian into text with most of the statistical features of Voynich A, including the slot grammar, and the author is careful to say that it reproduces the features of B less well and that it is a demonstration of possibility, not a decipherment (Greshko, 2025).

A verbose cipher has costs. It inflates the text—a page of Latin becomes two or three pages of symbols—and the Voynich manuscript, if it is a cipher of this kind, is a short book. It must also account for the line effects which an enciphering clerk working letter by letter would not produce unless the plaintext had been laid out line by line first. And it must account for the fact that the sections differ in vocabulary in the way real books do, which it does naturally if the plaintext does. The family survives, in this form, with the burden of explaining why the cipher has not been reversed given the quantity of text: thirty-seven thousand words is a great deal of ciphertext for a fifteenth-century system.

CONSTRUCTED SYSTEM

The third family holds that the text is

neither a natural language nor a cipher of one but something made: a philosophical language of the kind that seventeenth-century thinkers like Wilkins and Dalgarno designed, a shorthand, a system of glossolalia written down, a notation for something other than speech. Such systems are usually regular to a fault, and the regularity of Voynichese is what makes the idea attractive. Against it is the date. The great age of invented languages begins two centuries after the manuscript, and no fifteenth-century system of the kind is known, though the *ars combinatoria* of Ramon

Lull and the abbreviation systems of the law courts show that the century had the ingredients. Again it too is the vocabulary, which behaves like the vocabulary of a book about things rather than of a category system: words cluster by subject and recur across the pages about one subject. The family is not excluded, but nothing specific supports it, and every specific proposal within it has been a reading from the outside in.

MEANINGLESS

The last possibility is that the text means nothing: that it was made to look like writing, for money, for prestige, or for reasons we would not now understand, by a procedure that produced plausible symbols without encoding anything. Gordon

Rugg argued in 2004 that a table of syllables and a card with holes cut in it, the Cardan grille, could generate Voynich-like text quickly (Rugg, 2004), Andreas Schinner found in 2007 that the text has statistical features of a stochastic process (Schinner, 2007); Timm and Schinner proposed in 2020 the auto-copying procedure described in chapter 9, in which the scribe writes each new word by copying a word already on the page and altering it, and shows that it reproduces the near-repeat rate, the slot structure and much of the line behaviour (Ti[illegible] and Schl[illegible]ner 2020).

This family has one great strength. It is the only one that explains the repetition without needing a genre, and it explains the slot grammar as a by-product of copying. It also fits the fifteenth century: there was a market at court for books of secrets, and a book that looked like one would have sold.

Against it stand three findings. The first is the long-range order documented by Montemurro and Zanette and confirmed, with the scale held constant, by our contrasts: the words are distributed across the book as the words of a book about several things are (Montemurro and Zanette, 2013). A generator would have to be steered section by section to produce this, which is possible but is a further design decision. The second is the labels: words beside drawings that are drawn from the general vocabulary, are rarely repeated, lack the line-initial marks, and are longer and rarer on the containers. A generator producing labels would have to know that labels are different. The third is the distance profile of the near-repeats. If words were copied from nearby, the rate of near-twins should depend strongly on order. It depends on it a little: among the ten words before any given word, 27 per cent contain a twin in the written order and 24 per cent when the words of the page are shuffled (chapter 17). Most of the near-twin rate is a property of the vocabulary—many words differ from one another by a letter and survives any reordering. Copying contributes the small remainder at best.

None of these is accurate. The hoax hypothesis can be elaborated to meet each. But each elaboration adds a rule, and the procedure that results— generate words by slots, vary the vocabulary by section, treat labels differently, mark line starts and ends, do it consistently across five

scribes— its a complicated as a cipher and harder to teach.

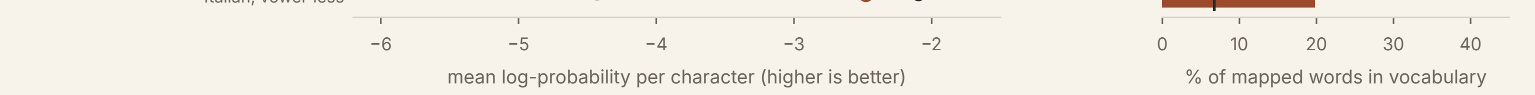
WHERE THIS LEAVES THE QUESTION

Two families are in, reasonable health: a natural languages in a non-alphabetic or positionally variant script, and a verbose cipher of a known language. One is alive but unsupported: a constructed system. One is alive but strained: meaninglessness. The two healthy families share a feature, which is the most useful thing the evidence has to say: in both, the symbols of the script are not letters in the ordinary sense. Several symbols together correspond to one unit of whatever is underneath, and the slots of chapter 6 are the seams between those units. That is where chapter 17 begins.

17

Attempting a Reading

THIS chapter reports what happened when we tried to read the text. We set out the ex-



Source: ZL3b running text; equal-length reference samples. Hill-climbing solver with 12 restarts; a control encipherment of each language was recovered exactly. Scores are mean log-probability per character (natural log).

Figure 17.1 The best one-to-one keys and what they produce.

periments in the order we ran them, with their results, including the negative ones, because a negative result with its method attached is worth more than a positive result without. Everything here can be re-run from the scripts in Appendix A in about a minute.

EXPERIMENT 1: THE BEST POSSIBLE KEY

The simplest cipher hypothesis is a one-to-one substitution: each Voynich glyph stands for one let-

Figure 17.2 Entropy at three levels of cutting.

Figure 17.3 Near-twins by distance.

ter of some language. Chapter 9 argued from the entropy that this cannot be right: We tested it directly anyway because the argument from entropy is indirect and because the test produces something the entropy does not: the best key there is, and what the text looks like under it. The method is standard. For a candidate language build a table of how often each letter follows each other letter in a sample of that language. Then search for the assignment of Voynich glyphs to letters that makes the Voynich text score highest under that table: start from a ran-done assignment, swap two letters, keep the swap if the score improves, repeat several thousand times, restart from scratch a dozen times, keep the best. The glyphs were the EVA units with the benches and gallows-benches as single symbols, the twenty-two to twenty-five commonest, mapped one-to-one onto the letters of the language. The word space was fixed to the word space.

The search works. As a control we took each reference text, enciphered it with a random substitution, and gave the result to the same search; in every case the search recovered a key whose score was identical to the plaintext's own, that is, it found the plaintext. Figure 17.1 shows what it found for Voynichese. For every language the best key scores far above a random key and below the language's own text. The gap to the language's own text is smallest for Pliny's Latin (0.07 nats per character) and largest for German (0.33). A gap of 0.07 nats per character sounds small; over 190,000 characters it is a likelihood ratio of $e^{12,000}$, which is to say that the Voynich text, under the best Latin key, is not Latin. What the best keys produce is shown in the right-hand panel of the figure. Under the best Latin key, 8 per cent of the mapped words are words that occur in the Vulgate, under a random key, under 1 per cent. Under the best English key, 7 per cent, German, 4; Italian, 9. The first line of the manuscript under the best Vulgate key reads *xuaed etus um ubullo pis pime hmrd e thm pisce dime gum*. The hit are short words that any assignment of common letters will produce by chance, *um*, *e*, *tim*, and they do not join up.

EXPERIMENT 2: WITHOUT VOWELS

Medieval scribes abbreviated heavily, and one family of proposals holds that the text is a language with its vowels

Attempting a Readdropped, as Hebrew and Arabic drop them. We repeat the search against vowelless versions of the five reference texts. The fit is worse, not better, the gaps widen to between 0.18 and 0.59 nats per character. The dictionary hit rates rise, to 19–38 per cent, but so do the random-key rates, to 6–

11 per cent, because a vowelless vocabulary is full of two- and three-consonant strings that almost any mapping will produce. The output is *?bdtg thbm bs, bvbllq frm frst*. We do not regard this as a test of the abjad hypothesis in general, only of its simplest form.

Chapter 16 concluded that in the surviving explanations the glyphs are not letters, and that several glyphs make up one underlying unit. If so, the right measure of the text's predictability is not glyph-by-glyph but unit-by-unit, and the question is whether, cut at the right joints, the text looks like language.

We cut it three ways (figure 17.2). As glyphs, the conditional entropy is 2.26 bits, far below the 3.1–3.4 of

the references. As slot-fillers—each word divided by the template of chapter 6 into its opening, gallows, middle, and ending, and each filler treated as one symbol—it is 2.64 bits, with 43 distinct fillers. That is closer but still below the natural range, which suggests either that the slots are not the units, or that the units are larger than single slots. As whole words it is 4.32 bits, and here the picture changes: the reference texts, cut into words, range from

3.0 to 4.4 At the level of the word, Voynichese is exactly as predictable as the Vulgate, the King James Bible or the Luther Bible.

That is a real result, and it should be stated carefully. It means that whatever process produced the sequence of words produced one with the same amount of order from word to word as a text in a language has. It does not mean the words are words of a language; a nomenclator, a syllabary with word-sized groups, or a well-steered generator could do the same. But it locates the level at which the text behaves like language: not the glyph, not the slot, but the word.

EXPERIMENT 4: IS ANYONE COPYING?

Timm and Schinner's hypothesis makes a prediction about distance: if a scribe wrote each word by copying and modifying a word already on the page, words should have their near-twins nearby. We measured, for every word on every page with thirty or-more words, whether an identical or one-letter different word occurred among the previous one, two, five, ten, twenty or fifty words, and compared the result with the same pages with their words shuffled (figure 17.3). The written order has more near-twins than the shuffled order at every distance, but not many more:

4 per cent against 3 at a distance of one word, 16 against 14 at five, 27 against 24 at ten, 53 against 52 at fifty. The Latin control, cut into 85-word pages, is first lower at every distance (0.4, 7, 13, 29 per cent). The near-twin rate is mostly a property of the vocabulary, present in any

order, the written order adds one to three percentage points. Something local is going on the adjacent repeats of chapters are read-

but; copying from the surrounding lines is not the main engine of the text's repetitiveness—and a procedure that would show a steeper curve.

EXPERIMENT 5: CRIBS

A crib is a piece of the known plaintext, and cribs are how most historical ciphers were actually broken. The manuscript appears to the

offer several. The month names on the zodiac pages are not cribs: they are in a later hand, in Latin letters, and attached to the pictures, not to the text. The labels are the real candidates. If the thirty labels around the Pisces wheel were the names of the thirty degrees of Pisces, or the names of thirty stars, in some language, one could look for a mapping that turned them into a known list. We tried the obvious structural test. Names of the degrees of one sign would share nothing in particular with the degrees of another; names of the stars would be unique. The 281 distinct zodiac labels include 37 that recur on other signs, four of them on four signs apiece, and they are built from the common vocabulary of the running text. Within a sign the labels do lean toward a shared openings; 71 per cent of the labels on the second Aries wheel (f. 71r) begin with **ot**, 52 per cent of the Gemini labels with **ok**, and the **ot**-heavy signs are the early ones, Pisces, Aries and Taurus, while the later signs favour **ok**. But **ot** and **ok** are the two com-modest openings of label words everywhere in the book, and no sign has an opening or ending of its own. The labels look, in short, like ordinary words of the text standing beside figures, with a drift in their spelling from one end of the zodiac to the other, not like a list of names. The pharmaceutical container labels look more like names: unique, long, and absent from the running text; but there are thirty-six of them, beside

containers that cannot be identified, which is not enough to fix a key.

EXPERIMENT 6: OTHER PEOPLE'S KEYS

We did not repeat the published decipherments, because each has already been examined and found wanting on the same grounds, which are worth stating once. Newbold's 1921 reading depended on marks that are ink cracks. The readings of the 2010s and 2020s that have reached the press—Stephen Bax's partial identifications of plant and star names, Nicholas

Gibbs's 2017 claim of abbreviated Latin, Gerard Cheshire's 2019 "proto-Romance," and several since—share a method: assign values to a few glyphs so that a few labels become words in a chosen language, then extend the values to a line of text and translate the result freely. Each produces readings that do not transfer the key that makes one label a plant name makes the next label nothing, and the running text under the same key is not

language in any sense that a stecked reader can verify (Bowern and Lindemann, 2021; Davis, 2017). Hauer and Kondrak's 2016 computational study, which treated the text as an anagrammed abjad and found Hebrew the best-scoring of many languages, is a different kind of work and its authors were careful about it; the fit was weak, the assumption of anagramming was unmotivated, and the decoded sample did not withstand a reading by Hebrew speakers (Hauer and Kondrak, 2016).

WHAT WE HAVE, AND WHAT WE DO NOT

We have not read the text, and we do not think anyone has. What the experiments add to the picture is this. No one-to-one key to a European language exists, with or without vowels; the best such keys produces no language. The text is as orderly as language at the level of the word and far more orderly than language at the level of the glyph, which places the seam between symbol and meaning at the word or just below it. The repetition is in the vocabulary, not the copying.

And the labels, which ought to be the cribs, are made of the same words as everything else.

If a reading is to come, we think it will come from the labels and the recipes together, with the sheets in their original order, and with a candidate system—a verbose cipher or a syllabary—whose units are word-sized. Chapter 19 says what such a reading would have to show.

18

A Defensible Reading

A sense Reading,
this in
book the

has used the word, is an account of what is on the page that can be defended line by line. This chapter gives ours. It is organised from the most secure statements to the least, and each paragraph says where its support comes from. Nothing in it is a translation.

WHAT THE BOOK IS

It is a single project. The parchment was prepared in one batch in the first third of the fifteenth century; the quires were numbered by one fifteenth-century hand; the five scribes share one script, one set of drawing conventions and one page design; the two languages are distributed across scribes and sections as if assigned. A book made over many years by unrelated people, or assembled from parts, would not look like this (Chapters 2, 4, 8.)

It was made by a group. Five hands is Davis's count, and the alteration of two of them sheet by sheet in the herbal shows that at least two worked on the same section at the same time, each to a plan. The sheets being out of order shows that the binder was not one of them (Chapters 4, 8.)

I

is organised by subject, as a compendium. Plants with texts; stars and the zodiac with labels; bathers with text; a cosmological diagram;

a pharmacy with labelled parts and containers; a long list of starred entries. These are the parts of a medical and natural philosophical compendium of the kind that fifteenth-century physicians and apothecaries owned, in which a herbal, a calendar, a treatise on regimen and a book of recipes were bound together.

No surviving compendium has these pictures, but every section has a genre to belong to. (Chapters 10–15.)

Its sequence is lost. The present order of the sheets is not the maker's, and the colouring was added, at least in part, after the order was lost. Any reading that depends on the order of pages is not yet possible (Chapter 4.)

WHAT THE TEXT IS

It is a consistent system. Thirty-seven thousand words follow the same rules of word formation, line formation and paragraph formation, across five scribes and two languages. The rules are not those of any known alphabet, and they were not improvised (Chapters 6, 7.)

Its units are not letters. The glyphs are too predictable from one another to be the letters of a language written out plainly; they are predictable in exactly the way the parts of a composite symbol are. The level at which the text has the orderliness of language is the word, or just below it. (Chapters 9, 17.)

Its two languages are two settings of one system. A and B differ in a handful of affixes and in the frequency of common words, not in their structure; the same scribe writes both subjects in the same language differently, and two scribes write the same subject in the same language alike (Chapter 8.)

It is about things, or made to look as if it were. The vocabulary of each section is its own, in proportion to the way real books on several subjects divide their vocabulary, and this holds with the scribe held constant. This is the strongest evidence that the text carries content, and it is evidence, not proof. (Chapters 8, 9.)

It is not a one-to-one substitution of Latin, Italian, German or English, with or without vowels. (Chapter 17.) It is not well explained as meaningless text generated by copying. The repetition that the copying hypothesis was built to explain is a property of the vocabulary, not of the sequence; the section-by-section vocabulary and the behaviour of labels require more of a generator than copying provides (Chapters 16, 17.)

WHAT THE LABELS ARE

They are words of the same system, without its line-marks. They are mostly unique, mostly **o-** initial, and mostly ordinary vocabulary. Thirty-seven zodiac labels recur on more than one sign, which argues against their being proper names of degrees or stars. The container labels in the pharmacy are the one class that behaves like names. (Chapters 7, 11, 14.)

WHAT THE PICTURES ARE

They are com-

perent, provincial, and conventional in kind without being copies of any known exemplar. The plants belong to the illustrated-herbal tradition; the zodiac emblems to the German tradition of about 1400–1500, with a crossbow Sagittarius; the bath to the balneological tradition, with a resemblance to the manuscripts of the *De balneis Puteolanis*; the fortification liners to the rosettes to northern Italy and the Alps. Taken together they place the book's makers in the German-speaking or northern Italian region in the generation after 1404, and no more precisely than that. (Chapters 10–13.)

WHAT WE THINK, AND HOW FIRMLY

Of the explanations that survive the evidence, we find the verbose-cipher family the most probable, by a modest margin, for three reasons: it explains the entropy, the word lengths and the slot structure as consequences of a single mechanism; it fits the century, in which substitution tables and nomenclators were the ordinary tools of secrecy; and a worked example now exists that reproduces most of the statistics of language A from Latin or Italian plaintext (Greshko, 2025). We hold this at the level of a working hypothesis, not a finding. The natural-language-in-a-syllabary family is close behind and cannot be separated from it by the tests we know how to run.

If the first hypothesis is right, then there is a plaintext, probably in Latin or a Romance or German vernacular, probably a medical compendium, and the key is a table that maps letters to glyph groups of the kind the slot grammar describes.

That table has resisted a century of attack, which means either that it is more complicated than the slot grammar suggests—with homophones, nulls and context rules—or that something about the plaintext, such as heavy abbreviation, makes the ciphertext short of the regularities that cryptanalysis needs. Both are possible, and both would be helped by the one thing nobody has yet had: the sheets in their original order.

WHAT WE DO NOT KNOW

We do not know the language, the author, the place, the purpose, or the meaning of a single word. We do not know whether the text is meaningful. We do not know the original order of the leaves, the original number of leaves, or when the book was first bound.

We do not know what the women in the water are doing, what the rosettes represent, or which plants the plants are. We do not know what the containers contain. We have set out what would change this in the next chapter.

19

Judging the Next Solution

SOMEONE will announce a decipherment of the Voynich Manuscript within a year of this book appearing, someone does most years. This chapter is for the reader who wants to know whether to believe it. The tests below are not ours. They are the tests that every successful decipherment of an unknown script has passed, from Champollion's hieroglyphs to

Ventris's Linear B, and that every failed reading of this manuscript has not. A claimed solution that meets them deserves attention. One that does not can be set aside without discourtesy, however distinguished its author.

THE TESTS

1. It is stated in advance and applied mechani-

cally. The key—the table from glyphs to letters, or words to meanings, or whatever it is—must be written down complete before any passage is read, and reading must consist of applying it. A key that is adjusted as one goes, or that leaves the reader free to choose between alternatives at each word, is not a key. Newbold’s reading failed here, so did every “partial decoding” of the past decade.

2. **It is applied by someone else.** A second person, given the key and a page the proposer has not read, must produce the same reading. If the proposer is the only one

who can make it work, it is not working.

3. It produces language. The output must be a text in an identifiable language that native readers or competent scholars of that language accept as grammatical and meaningful, without the proposer's gloss. "Proto-Romance" that no Roma[illegible] philologist recognises, or Latin that is a string of abbreviations each expanded as convenient, does not count.

4. It reads the running

text, not only labels. Nine tenths of the manuscript is paragraph text. A reading of a dozen labels is a hypothesis. about a dozen words. The test is a page of the recipe section, or a herbal paragraph, read in full.

5. It is consistent with the pictures. The herbal paragraphs should discuss plants; the recipe entries should contain ingredients that also appear in the pharmacy; the bathing text should be about what the bathing pictures show, whatever that is. A reading in which the page about a plant turns out to be about something unrelated has a problem.

6. **It explains the statistics.** The key, applied in reverse to a real text in the claimed language, should produce text with the properties measured in Part II: the slot structure, the entropy,

the line effects, the two languages. If the proposed system could not have produced the manuscript, it did not.

7. It is consistent with the object.

A reading that requires the book to be sixteenth-century, or Mexican, or the work of one person, has to overcome the radiocarbon date, the pigments, the five hands and the quire marks. These are not interpretations and they do not move.

8. It accounts for its own novelty. If the language is one that a century of experts failed to see, the reading should explain what they missed and why it is only visible now. “Nobody tried

X″ is sometimes true. It is rarely true of the obvious languages.

WHAT WOULD MOVE THE QUESTION

Short of a decipherment, several pieces of work would change what can be said, and most of them are within reach. *The order of the sheets.* Layfield and Davis have proposed to use the water stains to restraints.

cover the original sequence (Layfield and Davis, 2026), the sewing holes, the paint transfers and the picture across folios 78v and 81r are further com-

A reconstruction tested against the text's own vocabulary, continuity would be the first fully independent check on whether the text has content.

A single transliteration with uncertainty. The two main transliterations agree on three words in four. A consolidated version that records the disagreements as alternatives, rather than choosing, would let every statistic in this book be given with an error bar.

The glyph inten-

tory settled. Davis's digital palaeography has begun this (Davis, 2020). A published, image-linked inventory of distinct forms, with their variants, would remove the largest known source of disagreement between analyses.

A verbose-cipher search. Greshko's cipher shows that a table of the right kind can produce the statistics (Greshko, 2025). The reverse problem—searching the space of such tables for one that turns the manuscript into language—has not been attempted at scale. It is a large computation but a definable one.

The containers. Thirty-six unique long labels beside thirty-six drawn vessels are the best cribs in the book.

Identifying the vessels, or finding their models in the material culture of the period, is work for a historian of pharmacy that has not been done.

The comparanda. The bathers, the rosettes and the zodiac wheels have each been compared with a few manuscripts. Systematic comparison against the digitised corpus of fifteenth-century medical and astrological manuscripts, now large, might find the models the drawings were made from, and a model with text would be a crib of a different kind.

None of this requires reading the manuscript. All of it would make reading it more likely, and all of it is the kind of patient, checkable work that the history of the subject has had too little of. The book has waited six hundred years. It can wait for the order of its pages.

A

Data and Methods

This appendix describes the data behind every number in the book and the programs that produced them, so that anyone can repeat the work. The programs are short, written in Python 3 with NumPy and Matplotlib, and are distributed with the book's source. Running the whole set, including the figures, takes under two minutes on an ordinary laptop.

SOURCE TEXT

ZL3b-n.txt. The Zandbergen–

Landini transliteration of the manuscript in EVA, version 3b of 13 May 2025, in IVTFF format 2.0, distributed at `https://www.voynich.nu/data/`. It records 5,225 loci (lines of text, labels, circular and radial text) on 226 pages and panels, with page variables giving each page′s quire ($Q), position in quire

($P), illustration type ($I), Currier language ($L), Davis scribe ($H) and Currier hand ($C), and markers for the start (<%>) and end (<$>) of paragraphs,

n.txt. Takeshi Takahashi's transliteration in the same format, used only for the agreement measure in chapter 5.

CD2a-n.txt. The Currier–D'Imperio transliteration, used

LT2a-

only to confirm the correspondence between Currier alphabet and EVA.

READING THE FORMAT

The reader (`ivtff.py`) applies these rules to each locus.

In-line comments `<...>` are removed. High-ASCII codes `@nnn;` for rare glyphs are removed. Ligature braces `{}` are removed and their contents kept. Where the transcriber gives alternatives [a:b], the first is taken. Uncertain spaces (,) are treated as spaces. Lines continued with a trailing / are joined. Words containing ? (an unreadable character) are excluded from word counts; there are 140 in the running text.

“Running text” means loci of generic type P (paragraph text) only Labels are loci of type L; circular and radial text are C and R. All counts in the book say which they use

GLYPH UNITS

Statistics are given on the EVA characters and, where stated, on a "glyph-unit" alphabet in which the sequences `cth`, `ckh`, `cph`, `cfh`, `ch` and `sh` are counted as single symbols. No other merging is done; `ii` and `ee` are left as sequences.

REFERENCE TEXTS

Five samples of exactly 34,180 words were taken from public domain texts: the Latin Vulgate and the Bible in English (King James), German and Italian from the parallel Bible corpus distributed by C. Christodoulopoulos and M. Steedman (`https:`

`//github.com/christos-c/bible-corpus`), and Books 2–5 of Pliny's *Natural History* from the Latin Library as distributed by the Classical Language Toolkit. Each sample is a contiguous stretch beginning one-fifth of the way into the source (one twentieth for Pliny), lowercased, with punctuation and digits removed. The samples are chosen for availability and size, not as candidates for the manuscript's language.

MEASURES

Entropy. $h_1 = -\sum_i p_i \log_2 p_i$ over single symbols; $h_2 = -\sum_{ij} p_{ij} \log_2 p_{j|i}$ over adjacent pairs; the word space is included as a symbol in both. *Index of coincidence:* $\sum_i n_i(n_i - 1)/N(N-1)$, excluding the space. *Zipf slope* least-squares slope of log frequency on log rank over ranks 1–1000. *Keyness* signed log-likelihood G^2 for each word's counts in two groups, words with fewer than 20 occurrences combined excluded; the share of “significantly skewed” tokens is the share lying in words with $|G^2| > 10.83$ ($p < 0.001$). *Group similar-*

ity: cosine between frequency vectors over the 400 commonest words. *B-score:* per page, the share of words containing `ed`, ending in `eey`, or beginning with `qok`, minus the share beginning with `cho` or `sho`, equal to `daiin`, or ending in `ol` or `or`. *Near-twins:* two words are near-twins if their edit-distance is 1. *Substitution search:* hill-climbing over one-to-one assignments of the n common glyph units to the letters of the reference alphabet, scored by $\sum_{ij} C_{ij} \log P_{m(i)m(j)}$ where C is the Voynich bigram count matrix and P the reference bigram model with add-0.5 smoothing; 12

rest of 6,000 proposals each, with the word space fixed.

FIGURES

All data figures were drawn by a small library (`vizkit`) that lays out each figure as

starts stack of measured text and plot regions at the final printed width, so that nothing is rescaled in the page, and that checks every rendered figure for overlapping or clipped text and for type below 6 points before it is written. The categorical palette was validated for color vision deficiency. The folio images are the Beinecke Library's IIIF images at 2,000 pixels wide, and the large foldouts at full resolution.

FILES

ivtff.py
IVTFF 2.0 reader.

corpus.py
Corpus object; selections by locus type, section, language, scribe, page.

refcorpora.py
Reference samples.

a1_structure.py
Codex map, section and quire

summaries.

a2_stats.py
Basic statistics, Zipf, word length, entropy.

a3_wordstructure.py
Glyph positions, word rules, slot grammar.

a4_lines.py
Line and paragraph effects.

a5_groups.py
A/B keyness, B-score,

`a9_decipher.py`
group similarity. The experiments of chapter 17.

`a6_labels_repeats_agreement.py`
Labels, repetition, transcription agreement.

`build_figures.py`, `figs_variants.py`, `vizkit/`
The figures.

B

Folio Concordance

Every page and foldout panel in the ZL3b transliteration, in bound order, with its quire, the type of illustration on it (after the transliteration's page variables), Currier's language where he assigned one, Davis′s scribe, and the number of words of running text and of labels. Section codes: H herbal, A astronomical, Z zodiac, B bathing (biological), C cosmological, P pharmaceutical, S starred text, T text only.

page	q.	type	lang.	scribe	text	labels
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	20
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	A	1	[illegible]	0
[illegible]	[illegible]	H	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	[illegible]	A	1	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0

Folio Con-
cor-

page	q.	type	lang.	scribe	text	labels
page	q.	type	lang.	scribe	text	labels
f75r	18	B	B	3	382	3
f99v	1387	BH	BA	21	3598	00
f96r	1837	TBH	BA	51	23451	5100
f90v	1839	BP	BA	51	3758	05
f99r1	1939	AP	BA	21	24137	06
f1002	1939	AP	BA	21	12560	6418
f100v2	1939	CP	BA	21	23739	1834
f101v1	1939	AP	BA	21	30267	1640
f103r1	1939	AP	BA	21	34079	2948
f102r1	1939	AP	BA	21	25907	2544
f102r2	1939	AP	BA	21	35637	1644
f102v2	1939	CP	BA	21	34615	200
f105r1	1949	TAP	BA	21	34137	09
f103r2	1240	CS	BB	23	15073	00
f103v4	1240	CS	BB	23	4563	4700
f104r6	1240	TCS	BB	23	49008	00
f104v5	1240	TCS	BB	23	34094	00
f105r3	1240	CS	BB	23	21778	00
f105v2	1650	ZS	AB	43	10929	3400
f106r1	1650	ZS	AB	43	7031	2500
f106v	1550	ZS	AB	43	14334	1760
f107r	1550	ZS	AB	43	14087	2120
f107v1	1550	ZS	AB	43	14262	2150
f108r2	1550	ZS	AB	43	21093	3210
f108r3	1550	ZS	AB	43	16940	4150
f108v3	1550	ZS	AB	43	16632	3490
f110v2	1550	ZS	AB	43	6685	3100
f112r2	1550	ZS	AB	43	4928	2830
f113v2	1250	ZS	AB	43	6019	3000
f113v1	1250	ZS	AB	43	5769	3000
f113rv	1370	BS	AB	213	315801	00
f114r	1370	BS	AB	213	36996	3600
f114rv	1370	THS	BB	233	56374	900
f115v	1370	BS	BB	2–	39361	000
f115r1	1370	BS	BB	233	34018	1100
f116r2	1370	BS	BB	233	35295	1000
f116v2	1370	BT	B–	233	25550	702
f98v1	137	BI	B	23	2914	0

C

Statistical Tables

Values computed from the ZL3b transliteration (running text only, words containing an unreadable character removed) and from reference samples of 34,180 words each. Entropies in bits; the character stream includes the word space as a symbol.

C.1 BASIC STATISTICS

corpus	tokens	types	hapax	mean len.	s.d.	Zipf slope	h_1	h_2
Voynich (all running text)	34,180	7,466	5,266	5.10	1.94	-1.04	3.88	2.13
Voynich A	10,772	3,462	2,530	5.03	2.11	-0.99	3.85	2.19
Voynich B	22,869	4,941	3,427	5.14	1.84	-1.07	3.88	2.01
Latin (Vulgate)	34,180	6,817	4,084	5.30	2.68	-0.95	4.00	3.33
Latin (Pliny, Nat. Hist.)	34,180	12,578	8,976	5.97	2.71	-0.81	4.05	3.42
English (KJV)	34,180	2,351	933	4.08	1.90	-1.17	4.01	3.11
German (Bible)	34,180	3,575	1,634	4.54	2.10	-1.08	4.06	3.15
Italian (Bible)	34,180	4,314	2,286	4.45	2.48	-1.07	4.01	3.20

Voynich figures on the glyph-unit alphabet (ch, sh, cth, ckh, cph, cfh as single symbols): all text $h_1 = 3.87$, $h_2 = 2.26$; A $h_2 = 2.38$; B $h_2 = 2.11$.

C.2 LABELS

type	description	tokens	types	% q-initial	% o-initial	mean len.	% types in text
Lz	zodiac figure labels	345	281	0.3	67.2	5.66	51
Ln	nymph labels (biological)	64	60	3.1	54.7	5.09	80
Lf	plant-part labels (pharmaceutical)	202	191	0.5	51.0	5.75	52
Lc	container labels (pharmaceutical)	36	36	0.0	52.8	6.94	8
Ls	star labels	86	85	1.2	59.3	5.77	56
Lt	tube/tub labels	55	53	1.8	69.1	5.76	60
L0	other labels	202	133	2.5	22.8	3.52	58
La	astronomical labels	9	9	0.0	55.6	5.78	56
Lp	herb labels	5	5	0.0	20.0	5.20	60
P	running text (reference)	34180	7466	15.5	20.4	5.10	100

C.3 LINE POSITION

Share of words (per cent) by initial or final glyph and by position in the line; 4,034 lines with three or more words.

position	o	c	q	s	d	y	a	p	k	t	f	l
line-first word, initial glyph	13.4	4.1	13.0	14.7	15.4	14.6	0.4	9.3	2.6	9.7	0.8	1.7
line-middle word, initial glyph	21.1	21.3	16.9	12.2	7.1	3.4	5.5	0.5	3.8	1.9	0.3	4.2
line-last word, initial glyph	22.9	16.6	8.8	8.0	16.8	3.5	8.7	0.4	3.0	1.7	0.2	5.9

	y	n	l	r	s	m	o	d	g
line-first word, final glyph	32.4	20.0	14.2	20.4	3.7	0.4	6.0	1.5	0.0
line-middle word, final glyph	42.2	16.2	16.5	15.2	3.0	0.9	3.2	1.4	0.1
line-last word, final glyph	39.0	15.3	11.1	8.6	3.4	15.6	1.3	2.6	2.2

C.4 WORDS MOST CHARACTERISTIC OF A AND OF B

Counts and rate per 10,000

words; ranked by signed log-likelihood G^2.

word	A	B	/10k A	/10k B
daiin	456	296	423	129
chol	261	99	242	43
chor	165	27	153	12
cthy	88	5	82	2
sho	94	14	87	6
cthol	51	2	47	1
chy	97	33	90	14
s	161	98	149	43
shol	107	56	99	24
cthor	38	2	35	1
cho	55	16	51	7
shor	64	26	59	11
qotchy	48	13	45	6
ckhey	27	1	25	0
otchol	25	1	23	0
kchy	29	4	27	2
shy	56	28	52	12
dchy	21	1	19	0
dchor	21	1	19	0
chain	35	10	32	4
d	33	9	31	4
ckhol	19	1	18	0
kchor	19	1	18	0
dor	43	20	40	9
ckhy	28	7	26	3

word	A	B	/10k A	/10k B
chedy	1	486	1	213
shedy	2	419	2	183
qokeedy	0	303	0	132
qokedy	1	267	1	117
qokain	4	274	4	120
otedy	0	138	0	60
qokaiin	22	243	20	106
okedy	0	106	0	46
lchedy	1	115	1	50
okeedy	0	103	0	45
qotedy	0	87	0	38
qol	6	132	6	58
qoteedy	0	76	0	33
oteedy	1	86	1	38
qokal	16	168	15	73
qokeey	40	262	37	115
okain	8	125	7	55
sheedy	1	81	1	35
chdy	8	122	7	53
otain	3	90	3	39
keedy	0	62	0	27
al	24	183	22	80
otar	8	113	7	49
aiin	81	366	75	160
ol	99	419	92	183

C.5 SUBSTITUTION SEARCH

Mean log-probability per character (natural log) under each language's bigram model. In every case the control (a random encipherment of the language’s own text, solved by the same search) was recovered with a score equal to the own-text score. “Hits” is the share of mapped Voynich words found in the reference sample's vocabulary, under the best key and under a random key.

language	own text	best key	random key	hits, best	hits, random
Latin (Vulgate)	-2.307	-2.477	-4.744	7.9%	0.6%
Latin (Pliny, Nat. Hist.)	-2.370	-2.437	-4.552	8.2%	2.5%
English (KJV)	-2.154	-2.443	-4.505	6.9%	1.1%
German (Bible)	-2.184	-2.516	-5.735	3.8%	0.3%
Italian (Bible)	-2.221	-2.501	-5.363	8.8%	1.8%
Latin (Vulgate), vowel-less	-2.131	-2.426	-3.487	28.4%	8.8%
Latin (Pliny, Nat. Hist.), vowel-less	-2.222	-2.398	-3.603	38.1%	11.1%
English (KJV), vowel-less	-1.968	-2.559	-4.366	19.2%	5.9%
German (Bible), vowel-less	-2.046	-2.563	-4.138	20.5%	8.7%
Italian (Bible), vowel-less	-2.096	-2.477	-4.430	19.8%	6.7%

D

How This Book Was Made

This book was written by a human author working with an AI system, Claude, made by Anthropic. The reader is entitled to know what that means in practice, and what it does not.

WHAT EACH DID

The author set the question, the standard of evidence and the shape of the book, chose the format and the figure style from alternatives, and is responsible for every claim in it. The AI system did the computation: it fetched the transliterations and reference texts, wrote and ran the programs in Appendix A, built the figure library, drafted the chapters from the results and the published literature, and checked the draft against the data.

Where the draft said something the data did not support, it was changed or cut. Where the literature was uncertain, the uncertainty is in the text.

WHAT WENT WRONG FIRST

An earlier draft of this project, made in the spring of 2026, was built from the manuscript images alone, without a machine-readable transcription. The AI system of that draft produced “readings” of labels and words by describing what it saw in the pictures and from those readings a theory grew: names of zodiac signs, names of planets, a seven-planet pharmacology, a constructed language. When the readings were checked against the transliterations in the October 2026, none of them existed. The strings the system had reported were not in the manuscript in any transcription; several contained sequences the script cannot form. The earlier draft was set aside and this book was begun from the data.

We record this because it is the characteristic failure of the subject, and because an AI system is unusually good at committing it. A language model asked to read an un-

readable page will read it, fluently and with confidence, and the result will have every outward feature of a finding. The only protection is the one this book has tried to apply throughout: no number without a program that produces it, no reading without a transcription behind it, no claim about the object without a source that examined the object. Readers who find places where that standard slips are asked to say so.

REPRODUCIBILITY

The transliterations, the reference texts, the programs and the figure sources are distributed with the book. Every figure can be regenerated from them, and every number in the text can be traced to a table those programs write. The folio images are the Beinecke Library's

ow free
and to
are use.

Bibliography

Diego R. Amancio, Eduardo G. Altmann, Diego Rybski, Osvaldo N. Oliveira, and Luciano da F. Costa. Probing the statistical properties of unknown texts: application to the Voynich manuscript. *PLoS ONE*, 8 (7): e67310, 2013.

Beinecke Rare Book and Manuscript Library. Cipher manuscript (Voynich manuscript). Beinecke MS 408. Yale University Library Digital Collections, https://collections.library.yale.edu/catalog/2002046; catalogue description at https://pre1600.ms.beinecke.library.yale.edu/docs/pre1600.ms408.HTM, 2020. Accessed October 2026.

William Ralph Bennett. *Scientific and Engineering Problem-Solving with the*

Computer. Prentice-Hall, Englewood Cliffs, 1976. Chapter 4 computes character entropies for the Voynich text.

Claire L. Bowern and Luke Lindemann. The linguistics of the Voynich manuscript. *Annual Review of Linguistics*, 7: 285–308, 2021. doi: 10.1146/annurev-linguistics-011619-030613.

Raymond Clemens, editor. *The Voyn-*

ich Manuscript. Yale University Press, New Haven, 2016. With an introduction by Deborah Harkness and essays by René Zandbergen, Arnold Hunt, Jennifer Rampling, Alain Touwaide, Greg Hodgins, and Paula Zyats and colleagues.

Prescott H. Currier. Some important new statistical findings. In M. E. D'Imperio, editor, *New Research on the Voynich Manuscript: Proceedings of a Seminar, 30 November 1976*. Washington, D.C., 1976. Reprinted in D'Imperio (1978) and at voynich.nu.

Lisa Fagin Davis. Voynich manuscript "solution" debunked. Statement on behalf of the Medieval Academy of America, responding to Nicholas Gibbs, *Times*

Literary Supplement, 5 September 2017, 2017.

Lisa Fagin Davis. How many glyphs and how many scribes? Digital paleography and the Voynich manuscript. *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies*, 5 (1): 164–180, 2020.

Lisa Fagin Davis. Voynich codicology. Manuscript Road Trip (blog), `https://manuscriptroadtrip.wordpress.com/2025/01/19/voynich-codicology/`, January 2025.

M. E. D'Imperio. *The Voynich Manuscript: An Elegant Enigma.* National Security Agency, Center for Cryptologic History, Fort George G. Meade, 1978.

Michael A. Greshko. The Naibbe cipher: a substitution cipher that encrypts Latin and Italian as Voynich manuscript-like ciphertext. *Cryptologia*, 2025. doi: 10.1080/01611194.2025.2566408. Published online 26 November 2025.

Bradley Hauer and Grzegorz Kondrak. Decoding anagrammed texts written in an unknown language and script. *Transactions of the Association for Computa-*

tional Linguistics, 4: 75–86, 2016.

Greg Hodgins. Forensic investigations of the Voynich manuscript. In Raymond Clemens, editor, *The Voynich Manuscript*. Yale University Press, New Haven, 2016. Radiocarbon dating carried out at the University of Arizona, 2009.

Arnold Hunt. Voynich

the buyer. In Raymond Clemens, editor, *The Voynich Manuscript*. Yale University Press, New Haven, 2016.

Gabriel Landini. Evidence of linguistic structure in the Voynich manuscript using spectral analysis. *Cryptologia*, 25(4):275–295, 2001.

Gabriel Landini and René Zandbergen. The European Voynich Alphabet (EVA). `https://www.voynich.nu/transcr.html`, 1998.

Colin Layfield and Lisa Fagin Davis. Singular structure and the Voynich manuscript. *Digital Medievalist*, 19, 2026. doi: 10.4000/16kcn. Published 19 October 2026.

Luke Lindemann and Claire Bowern. Character entropy in modern and historical texts: comparison metrics for an undeciphered manuscript. arXiv:2010.14697, 2021.

and co-occurrence patterns in the Voynich manuscript: an information-theoretic analysis. *PLoS ONE*, 8(6):e66344, 2013.

Marcelo A. Montemurro and Damián H. Zanette. Keywords

Hugh O'Neill. Botanical observations

the Voynich MS. *Speculum*, 19 (1): 126, 1944.

Nicholas Pelling. *The Curse of the Voynich: The Secret History of the World's Most Mysterious Manuscript*. Compelling Press, Surbiton, 2006.

Gordon Rugg. An elegant hoax? A possible solution to the Voynich manuscript. *Cryptologia*, 28 (1): 31–46, 2004.

Andreas Schinner. The Voynich manuscript: evidence of the hoax hypothesis. *Cryptologia*, 31(2): 95–107, 2007.

Jorge Stolfi. A prefix midfix suffix decomposition of Voynichese words; the core–mantle–crust word paradigm. https://www.ic.unicamp.br/~stolfi/voynich/, 1997–2000.

John H. Tiltman. The *Voyn-*

ich manuscript: “the most mysterious manuscript in the world”. *NSA Technical Journal*, 12(3):41–85, 1967. Reporting work begun in 1951; declassified.

Torsten Timm and Andreas Schinner. A possible generating algorithm of the Voynich manuscript. *Cryptologia*, 44(1):1–19, 2020.

René Zandbergen. The Voynich manuscript. https://www.voynich.nu, 1996–2026. Including the pages on the manuscript's description, the quire structure and page order, the illustrations, the transliteration files (ZL3b, 13 May 2025) and the IVTFF format definition. Accessed October 2026.

René

Zandbergen. Earliest owners. In Raymond Clemens, editor, *The Voynich Manuscript*. Yale University Press, New Haven, 2016.

Paula Zyats, Erin Mysak,

Jens Stenger, Marie-France Lemay, Anikó Bezur, and David D. Driscoll. Physical findings. In Raymond Clemens, editor, *The Voynich Manuscript*. Yale University Press, New Haven, 2016.

Summarizing the materials analysis, including that of McCrone Associates (2009).