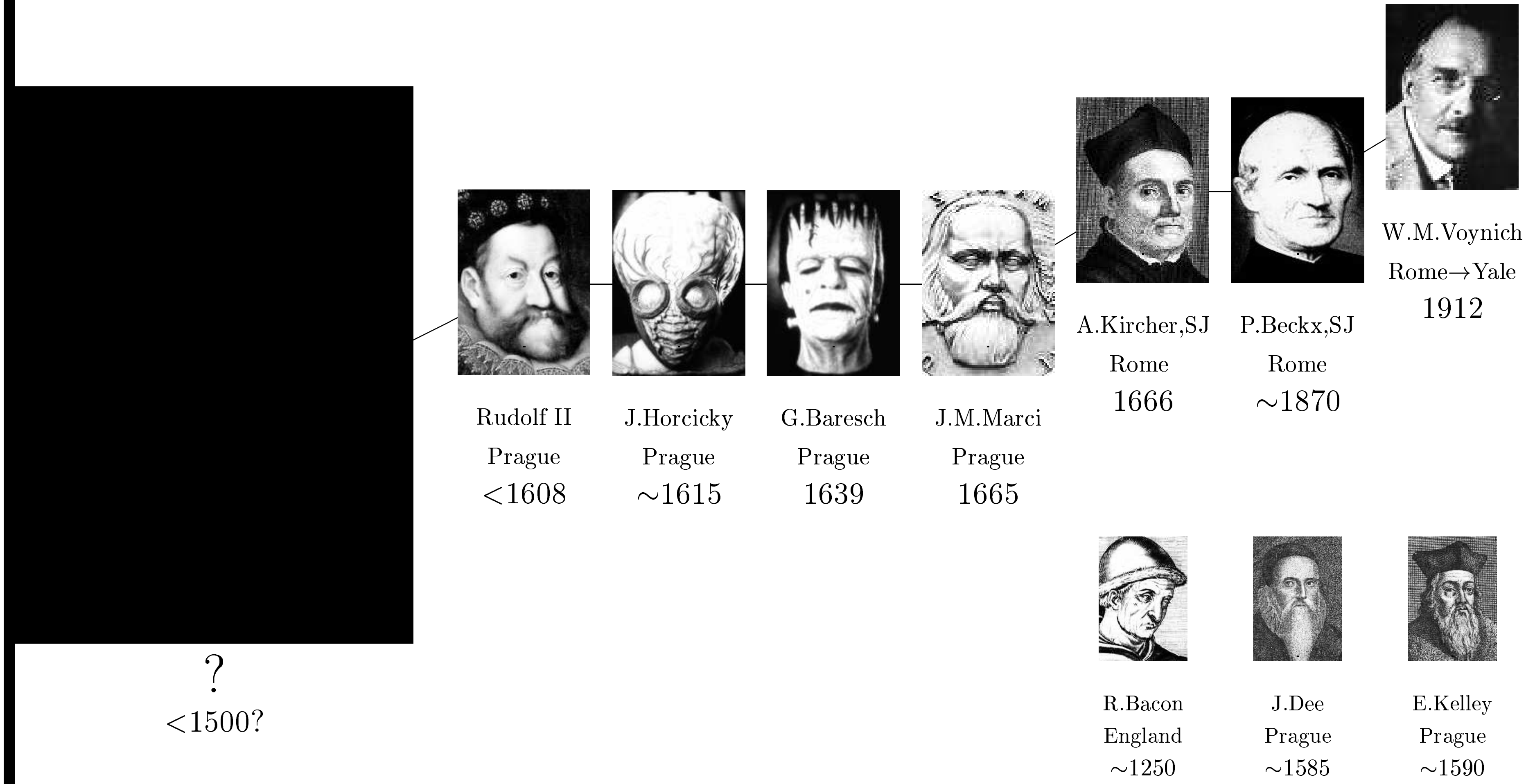
The Voynich Manuscript

Mathematical and Combinatorial Aspects

Jorge Stolfi

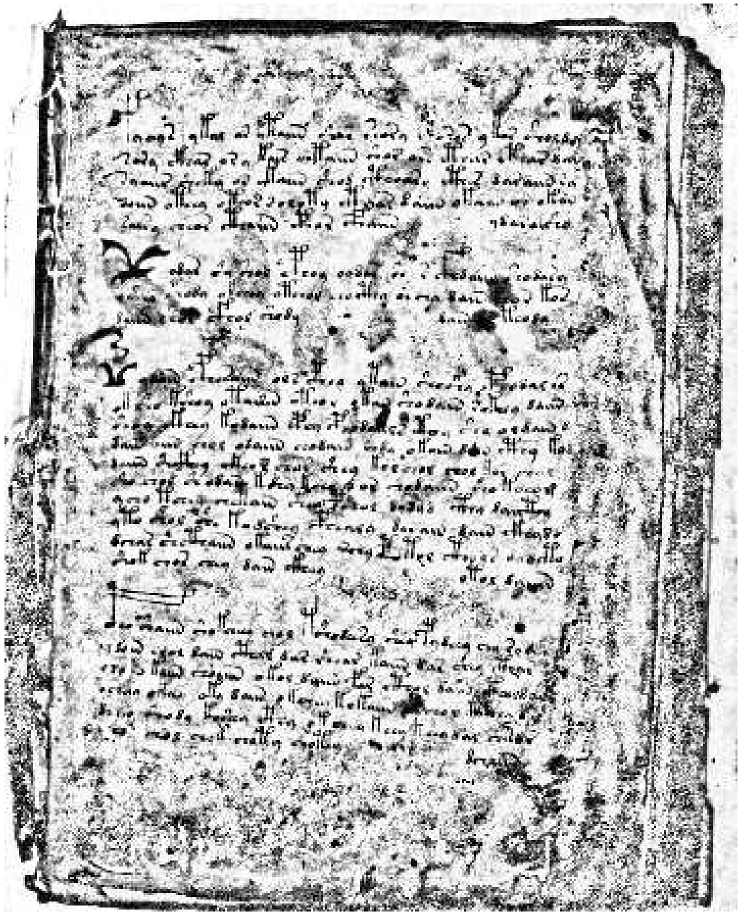
Institute of Computing
University of Campinas
(UNICAMP)

History of the Manuscript



images: Rene Zandbergen and other sources.

The Manuscript



Page f1r

location: Beinecke Library, Yale Univ.

name: MS 408; Voynich Manuscript (VMS).

dimensions: 15×22 cm.

length: 243 pages, ~40000 words.

illustrations: ~200.

material: vellum (calf skin), quill.

date: ~1450-1550?.

provenance: European?

subject: medicine? astrology?

title: unknown.

author: unknown.

language: unknown.

alphabet: unknown.

page images: Marshall Library

Major research efforts (inconclusive)

- 1930 J. Manly (WW I cryptographer).
- 1930–60 Father T. Petersen.
- 1930–60 P. Currier (US Navy crypt.).
- 1944–46 *First VMS Study Group*: W. F. Friedman, M. D'Imperio, …(NSA).
- 1950–60 J. Tiltman (UK crypt.)
- 1962–76 *Second VMS Study Group*: W. F. Friedman, M. D'Imperio, P. Currier, …(NSA).
- 1991- *Voynich Mailing List*: J. Gillogly, J. Reeds, J.Guy, G. Landini, R. Zandbergen, ….

source: Jim Reeds

False solutions

- 1928 W. Newbold (Roger Bacon's).
- 1943 J. M. Feely (Roger Bacon's).
- 1948 L. C. Strong (Anthony Ascham's).
- 1977 R. Brumbaugh (Dee/Kelly swindle on Rudolph II).
- 1978 J. Stojko (letters in Old Ukranian).
- 1987 L. Levitov (Cathar's rites in mixed German/French).

(all dates are approximate).

Contents

“herbal”
126 pages

Page f16v

“astronomical”
16 pages

Page f67r2

“biological”
20 pages

Page f77v

“pharmaceutical”
16 pages

Page f88r

“zodiac”
12 pages

Page f72v2

“starred”
25 pages

Page f105r

“other”
9 pages

Page f66r

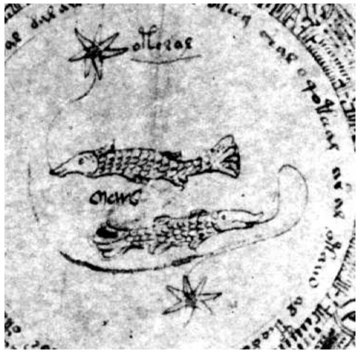
The illustrations

Page f5v

Page f25v

Page f81r

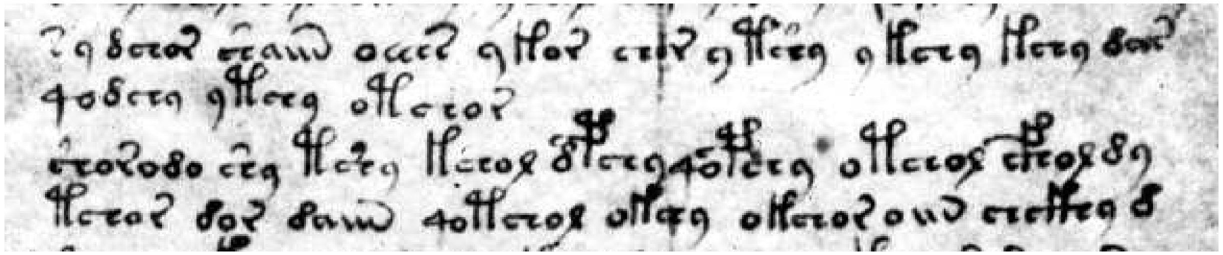
Page f83r



Page f70v2

- Plant drawings not identifiable (copied? invented?).
- Meaning of other drawings mostly unknown.
- Hundreds of women (almost all naked); only 2-3 men (dressed).
- Hairstyles, hats, and dresses are European, 1480-1520 [J. Porter].
- Crude "provincial" style, no planning, no compass or square.
- Almost no identifiable symbols or symbology.
- Inconsistent perspective.
- Colors may have been added later.

The main text



page f13r

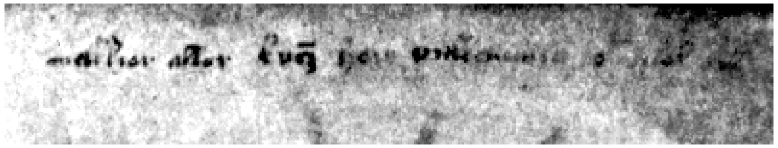
- No punctuation.
- No identifiable numerals.
- No identifiable symbols.
- No readable words (Roman, Greek, Hebrew...).

The round, upright letter style resembles the "humanistic hand" (Europe, late to mid 1400's); displaced by the compressed and slanted "italic hand".

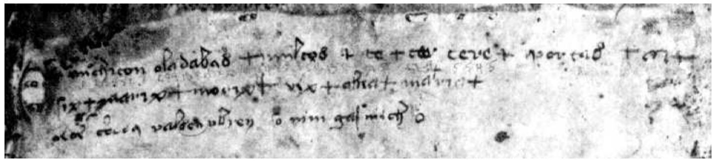
Readable bits of text

pages f70v2–f73v
month names in Zodiac

page f66r,
bottom margin



page f17r, top margin



page f116v (back cover)

- Some (all?) annotations may be apocryphal.
- The handwriting is very poor (illiterate?).
- **The language cannot be identified.**

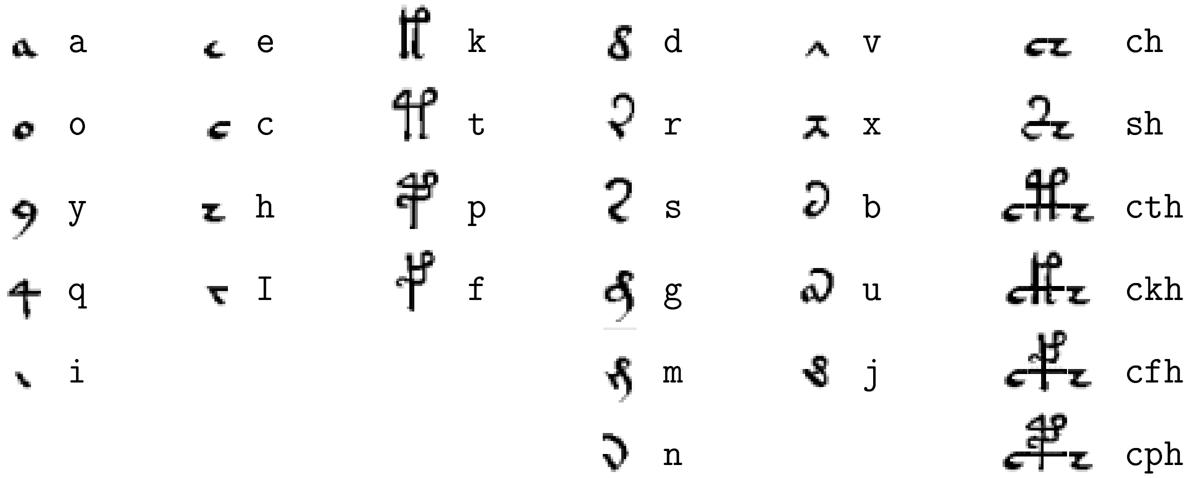
What is the alphabet?

Page f57v

Some hints on letter boundaries:

- Short words.
- Letter sequences.
- Allowable combinations.
- Interruption by figures.

The EVA transcription alphabet



Publicly available machine transcriptions

Partial computer transcriptions by First Study Group (~1950)
Recovery and additions by J. Reeds, J. Guy, J. Grove, and others (~1995)
Complete transcription by T. Takahashi (1998)
Verified transcription by G. Landini and R. Zandbergen (1999?)

font: Gabriel Landini

What does it mean?

The nature of the text is still unresoved:

- Cypher.
- Hoax/fraud.
- “Alchemical herbal.”
- Glossolalia.
- Musical notation.
- “Ignorant scribe” copying foreign manuscript.
- Invented language.
- Exotic natural language.
- ?

The case for cipher

Cryptography became fashionable around that time, e.g:

- Tranchedino (Milan, late 1400's): catalog of cipher alphabets— simple substitution, some nulls and digraphs.
- Cornelius Agrippa (Nettesheim, ~1500): many-to-one encoding, decoding relies on sense.
- Abbot Trittheim (Wurzburg, ~1500): "Trithemian cipher," one-to-many encodings, hiding message among lots of noise.

The case against cipher

- Word frequencies seem natural.
- Distribution of words over pages seems natural.
- Labels occur in nearby text.
- Cryptological attacks have failed so far.
- Why encode it?

Vigenère cipher

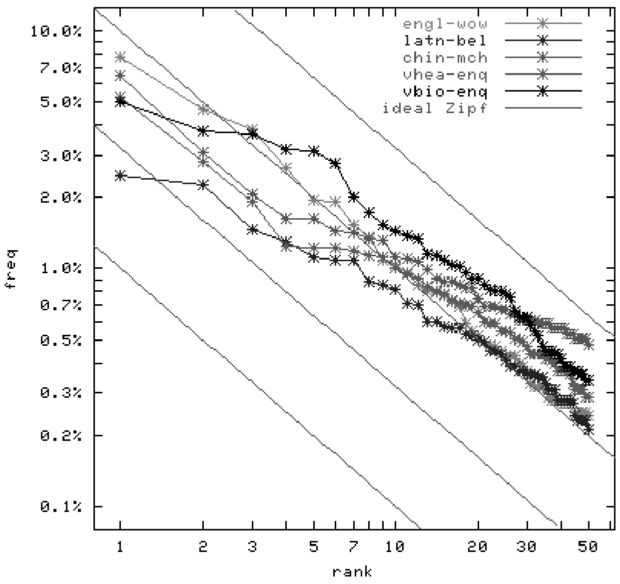
Typical ciphers of the period were variations on the Vigenère method.

Simple Vigenère cipher with key `012345`:

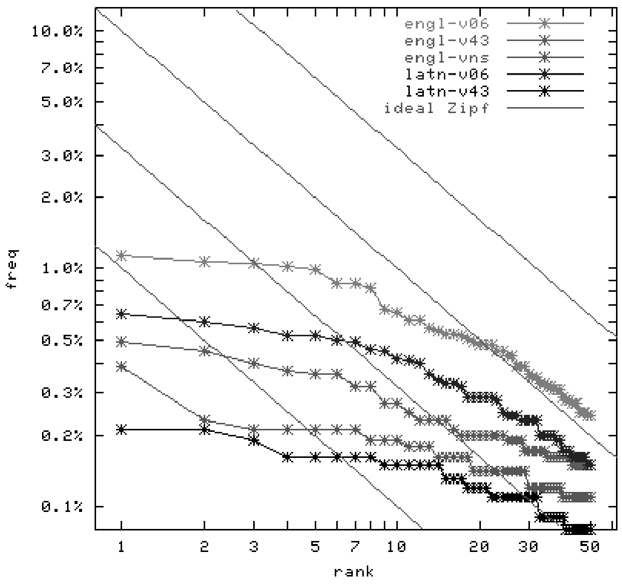
```
Plaintext:     the bird saw the cat and the dog
Repeated key:  012 3450 123 450 123 450 123 450
Ciphertext:    tig emwd tcz wle dcv esd ujh gtg
               ---          ---         ---
```

Key of length $m \Rightarrow$ each word gets mapped to $\sim m$ words, with equal probability.

Effect of Vignère cipher on Zipf's law



Frequency vs. rank
Plaintexts



Frequency vs. rank
Ciphertexts

Is it random text?

Simple two-state Markov model that avoids words of length 0:

$$\textbf{state } \boxed{\mathbf{0}} \textbf{ :} \quad \text{emit letter and go to } \boxed{\mathbf{1}}\,.$$

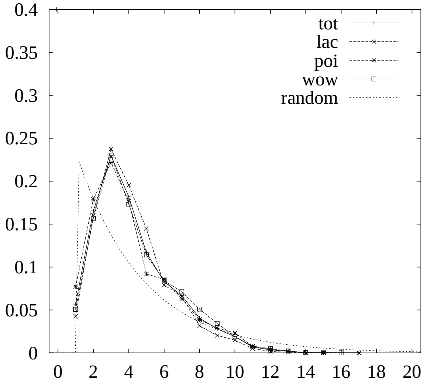
$$\textbf{state } \boxed{\mathbf{1}} \textbf{ :} \quad \begin{cases} \text{with prob. } p\text{, emit letter and stay in } \boxed{\mathbf{1}}\,; \\ \text{with prob. } 1-p\text{, emit space and go to } \boxed{\mathbf{0}}\,. \end{cases}$$

Average word length: $1/(1-p)$.

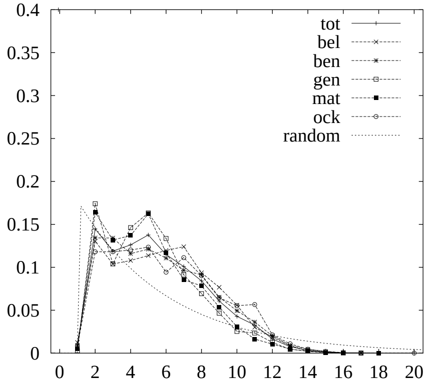
Frequency in text of words of length m: $(1-p)\,p^m$

Word length distribution

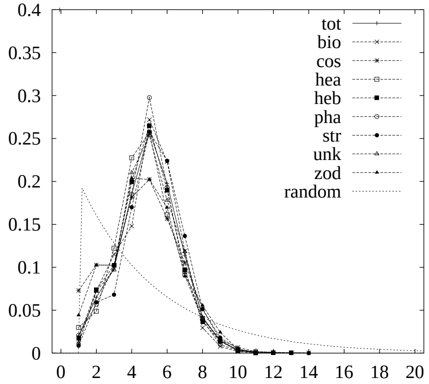
Word frequency in text as function of word length



English



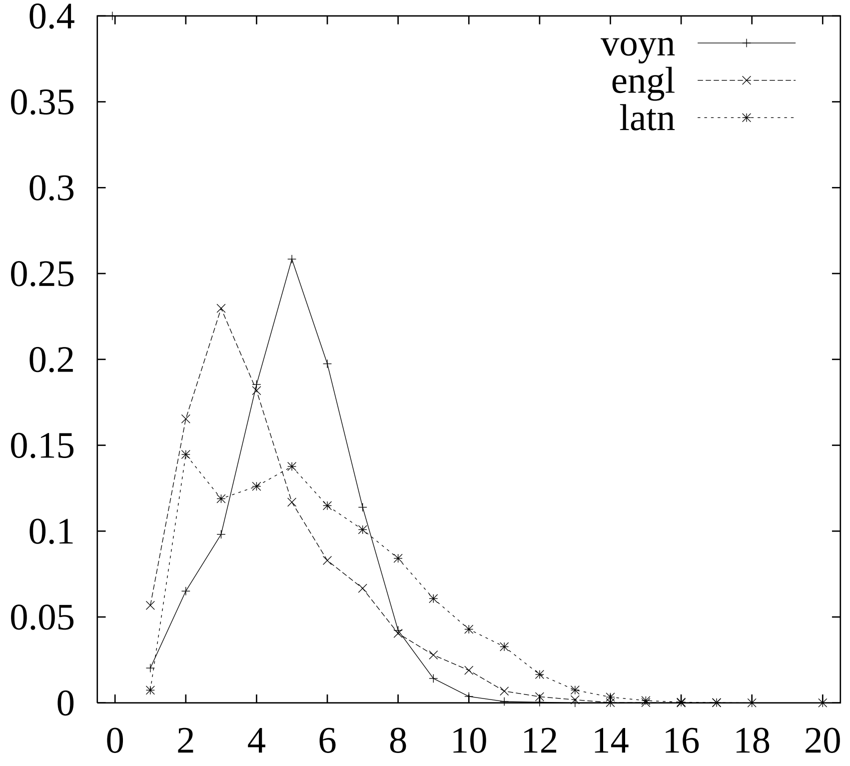
Latin



Voynichese

after Gabriel Landini

Comparing word length distributions



after Gabriel Landini

Entropy of Voynichese

Briefly:

- Information contents per word: $\sim$10 bits (similar to Latin and English)
- Information contents per character: depends on encoding (but more efficient than English and Latin).

Two languages? (1)

P. Currier ($\sim$1970) noted two "languages" (A and B):

- Different word frequencies.
- Different digraph frequencies.
- Perhaps different handwritings?
- Mostly split by section boundaries.
- Intermixed in the herbal section.

R. Zandbergen, G. Landini, and others (1998) confirmed with correlation and clustering methods.

However, the differences may be due to different subject matter or tense.

There may be a gradual variation; discontinuities may be due to lost pages.

Two languages? (2)

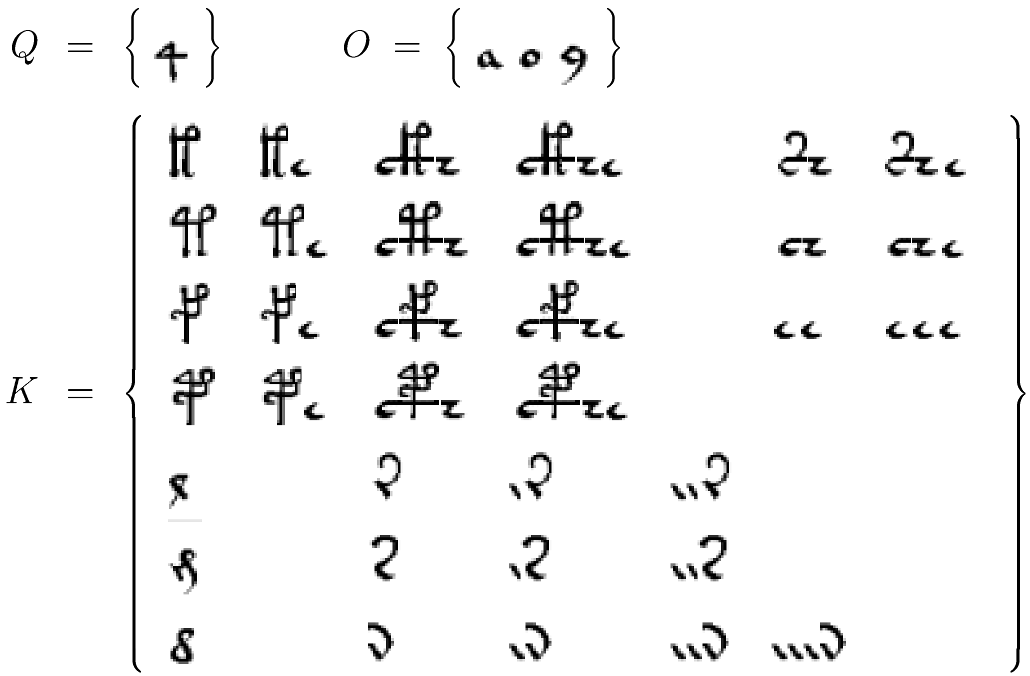
% Pharma	% Herbal-A	% Herbal-B	% Stars	% Astro	% Zodiac	% Biol
4.4 daiin	5.0 daiin	2.2 daiin	1.7 aiin	2.0 ar	2.8 al	3.6 shedy
1.9 chol	2.7 chol	1.9 aiin	1.6 chedy	1.7 aiin	2.7 aiin	3.5 ol
1.5 cheol	1.8 chor	1.9 or	1.4 ar	1.5 or	2.1 ar	3.1 chedy
1.5 ol	1.5 s	1.9 chedy	1.4 qokeey	1.4 daiin	1.4 oteey	2.3 qokedy
1.4 or	1.3 dy	1.5 chdy	1.2 al	1.4 dar	0.8 daiin	2.2 qokeedy
1.4 aiin	1.2 shol	1.4 dar	1.2 daiin	1.3 ol	0.8 dy	2.1 qokain
1.4 dal	1.2 cthy	1.3 ar	1.1 qokeedy	1.2 al	0.8 otaiin	1.7 qol
1.2 s	1.2 chy	1.1 dy	1.0 chey	1.0 o	0.8 oteody	1.6 qokal
1.1 okeol	1.1 sho	1.1 qokedy	1.0 qokaiin	1.0 r	0.8 s	1.4 shey
1.0 okeey	0.9 dain	1.0 chckhy	1.0 shedy	0.9 dal	0.7 dal	1.3 chey
0.9 qokeol	0.8 dar	1.0 shedy	0.9 ol	0.9 y	0.7 dar	1.2 qokaiin
0.9 cheor	0.8 shor	0.9 okaiin	0.8 okaiin	0.8 dy	0.7 o	1.2 qokeey
0.9 chey	0.7 shy	0.9 ol	0.8 okeey	0.7 oteey	0.7 oteos	1.1 daiin
0.9 chor	0.7 ol	0.9 okar	0.7 shey	0.7 s	0.6 otey	1.0 dar
0.9 dar	0.7 or	0.8 qokar	0.6 otaiin	0.6 chol	0.6 y	0.9 dal
0.8 qokeey	0.6 chey	0.7 okedy	0.6 chol	0.6 dair	0.5 air	0.9 or
0.8 qokol	0.6 cthol	0.7 dal	0.6 cheey	0.5 l	0.5 chey	0.8 qoky
0.8 shey	0.6 qotchy	0.7 okal	0.5 or	0.5 am	0.5 otar	0.8 lchedy
0.7 dol	0.5 dal	0.6 s	0.5 otal	0.5 chey	0.4 aly	0.7 dy
0.7 cheey	0.5 dol	0.6 cheky	0.5 qokain	0.4 air	0.4 am	0.7 otedy
0.7 cheody	0.5 cthor	0.6 otedy	0.5 otedy	0.4 chedy	0.4 cheey	0.6 qotedy

The fine structure of words

Denote $X^? = X$ or empty, $X^* =$ zero or more Xs. Then

typical word $= Q^?O^?(KO^?)^* = Q^?O^?KO^?KO^? \cdots KO^?$ (> 98% of text)

where



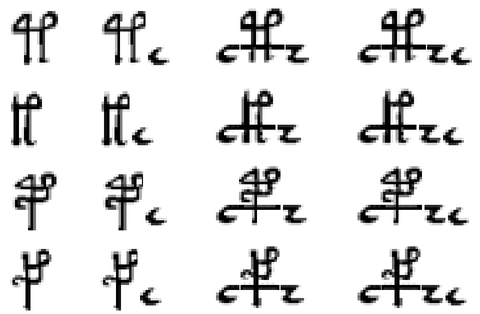
The coarse structure of words

Ignoring $\{$  $\}$:

$$\text{typical word} \;=\; Q^? R^? X^\alpha H^? X^\beta R^\gamma F^? \qquad (> 97\% \text{ of text})$$

where $\alpha + \beta$ and γ are 0, 1, or 2; and

$Q \;=\; \{$  $\}$ (prefix)

$H \;=\; \{$  $\}$ (core)

$X \;=\; \{$  $\}$ (mantle)

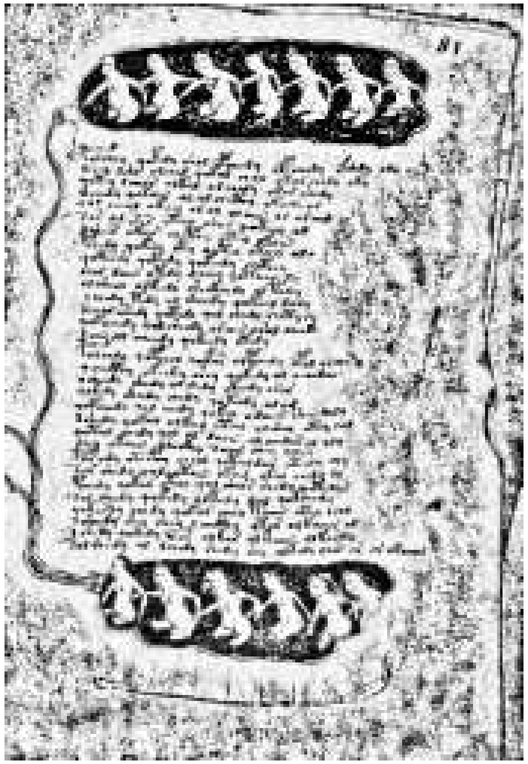
$R \;=\; \{$  $\}$ (crust)

$F \;=\; \{$  $\}$ (suffix)

Problem: The original page order

We know that the pages were bound out of order:

page f78v



page f81r

How can we discover the original page order?

Comparing pages by word usage

Assumption: a **single running text** continuing across page boundaries.

A topic-specific word will tend to occur in a few clusters of consecutive pages.

Let $f_{p,w}$ be the frequency of a given word w in page p, where $p = 1, 2 \ldots P$.

Find a permutation π of the pages that minimizes the change in the frequency of w from one page to the next:

$$S_w^2(\pi) = \sum_{i=2}^{P} (f_{\pi(i),w} - f_{\pi(i-1),w})^2$$

Trivial solution: sort all pages according to the frequency of w.

Comparing frequencies of several words

We want to minimize $S_w^2(\pi)$ simultaneously for several words $w = 1, 2 \ldots, W$.

$$\begin{aligned} S^2(\pi) &= \sum_{w=1}^{W} \sum_{i=2}^{P} (f_{\pi(i),w} - f_{\pi(i-1),w})^2 \\ &= \sum_{i=2}^{P} \sum_{w=1}^{W} (f_{\pi(i),w} - f_{\pi(i-1),w})^2 \end{aligned}$$

For each page p, consider the point $f_p = (f_{p,1}, f_{p_2}\ ldots, f_{p,W})$ of $\mathbf{R}^W$.

$$S^2(\pi) = \sum_{i=2}^{P} \mathrm{dist}^2(f_{\pi(i)}, f_{\pi(i-1)})$$

where $\mathrm{dist}^2(,)$ is the square of the Euclidean distance.

This is a variant of the Traveling Salesman problem (note that $\mathrm{dist}^2(,)$ does not satisfy the triangle inequality).

TSP by exhaustive enumeration

Approximate algorithms exist but aren't good enough.

Can we use exhaustive enumeration?

The Biological section has 20 pages.

$20! = 2,432,902,008,176,640,000\ldots.$

But not all page orders are possible!

- **quire**: a set of nested *bifolios*.
- **bifolio**: a sheet of vellum, folded in two *folios*.
- **folio**: a leaf of the bound book, with two *pages*.

How many ways are there to bind n bifolios?

Folio permutations

$$B_n = 2^n \, n! \, C_n$$

where

$$C_n = \text{number of balanced strings of } 2n \text{ parentheses}$$

$$()()()() \qquad (())()() \qquad (()())() \qquad \cdots \qquad (((())))$$

$$C_n = \binom{2n}{n} \frac{1}{n+1}$$

n	C_n	B_n
0	1	1
1	1	2
2	2	16
3	5	240
4	14	5376
5	**42**	**161,280**
6	132	6,082,560

Enumeration and testing can be done in about 15 minutes by an `awk` program!

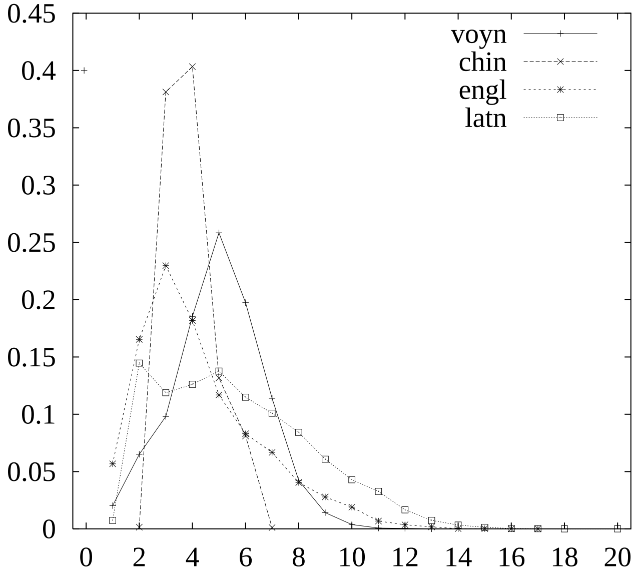
Other problems

- Ordering of the sections
- Spelling variation
- Grammatical structure

The Pizza Bet: phonetic Chinese?

```
dao4 ke3 dao4 fei1 chang2 dao4
ming2 ke3 ming2 fei1 chang2 ming2
wu2 ming2 tian1 di4 zhi1 shi3
you3 ming2 wan4 wu4 zhi1 mu3
```

- Short words only.
- Rigid word structure.
- No punctuation.
- Requires special alphabet.
- Spelling varies with context.
- Numbers are spelled out.
- Hidden grammar.
- No Western symbols.
- Various dialects.
- Possible in [1580..1600]



Word frequency by length

Some references

- *The Voynich "Roger Bacon" Cipher Manuscript.* MS 408, Beinecke Rare Book Library, Yale University.
- The members of the *Voynich Mailing List* (`voynich@rand.org`), mainly R. Zandbergen, J. Reeds, and Jim Gillogly.
- Jim Reeds, *Voynich Manuscript Bibliography* `http://www.research.att.com/ reeds/voynich/bib.html`
- Mary D'Imperio, *The Voynich Manuscript–An Elegant Enigma.* Aegean Park Press, ∼1978.
- David Kahn, *The Codebreakers.* Macmillan, 1967.

Image sources

- Roger Bacon's Portrait: MacTutor History of Mathematics archive
  `http://www-groups.cs.st-and.ac.uk/history/PictDisplay/Bacon.html`
- Portraits of P. Beckx, A. Kircher, J. M. Marci, Mondragone villa, Rudolf II: Rene Zandbergen's VMS site
  `http://www.geocities.com/Athens/Delphi/8956/`
- Voynichese alphabet: Gabriel Landini's site