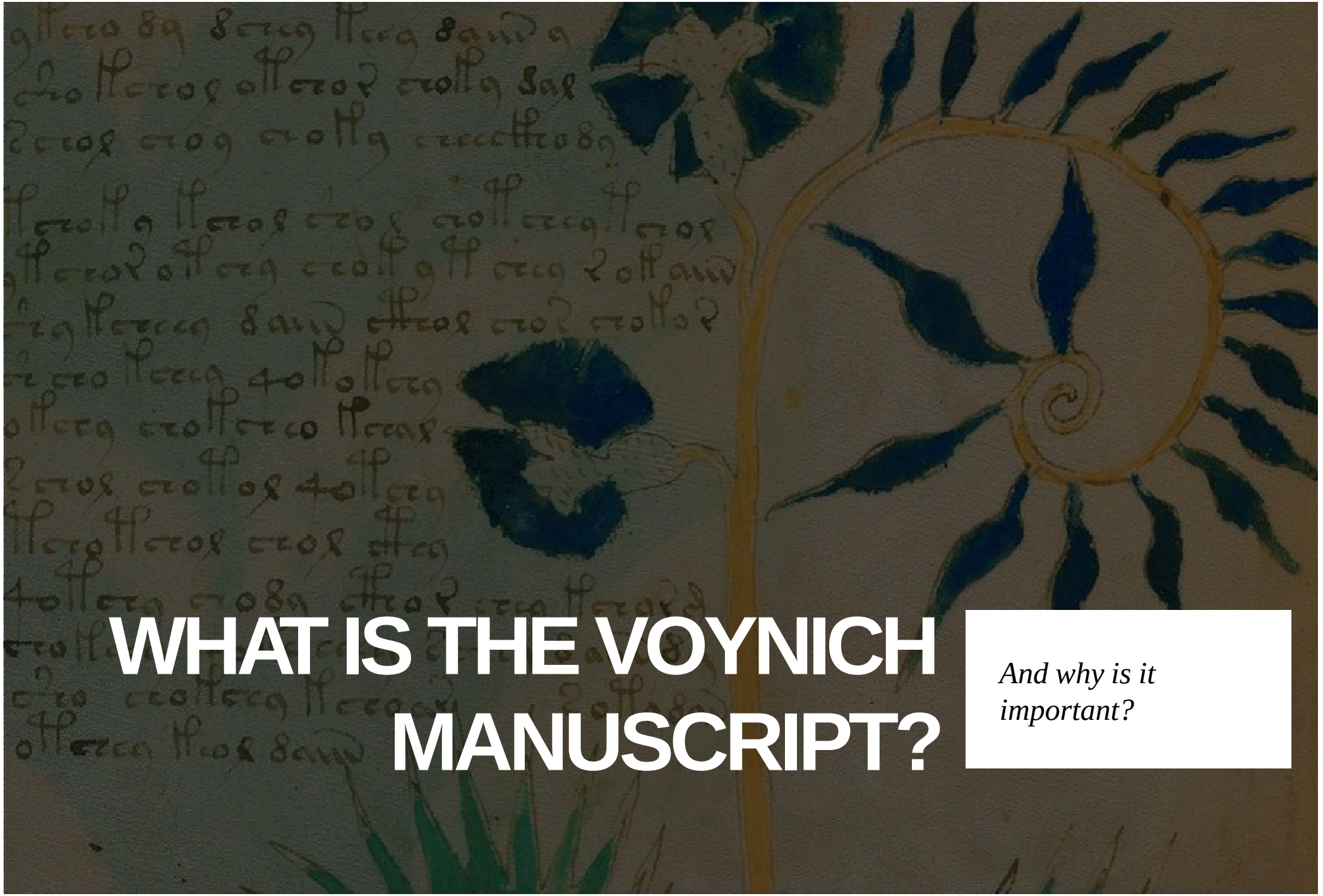
ON IDENTIFYING NATURAL LANGUAGES

A case study on the Voynich Manuscript

João Figueira



WHAT IS THE VOYNICH MANUSCRIPT?

And why is it important?

WHAT IS THE VOYNICH MANUSCRIPT?

A 15th Century Illustrated Book

- Acquired by Polish book seller Wilfrid Voynich (Michał Habdank-Wojnicz) in Italy, in 1912
- 200 pages of vellum (goat skin)
- Text on every page and big illustrations on most of them
- About late medieval herbology, astrology, pharmacology, and health
- Currently at the Yale library of rare books

WHY IS IT INTERESTING?

A Real-Life Indiana Jones Mystery

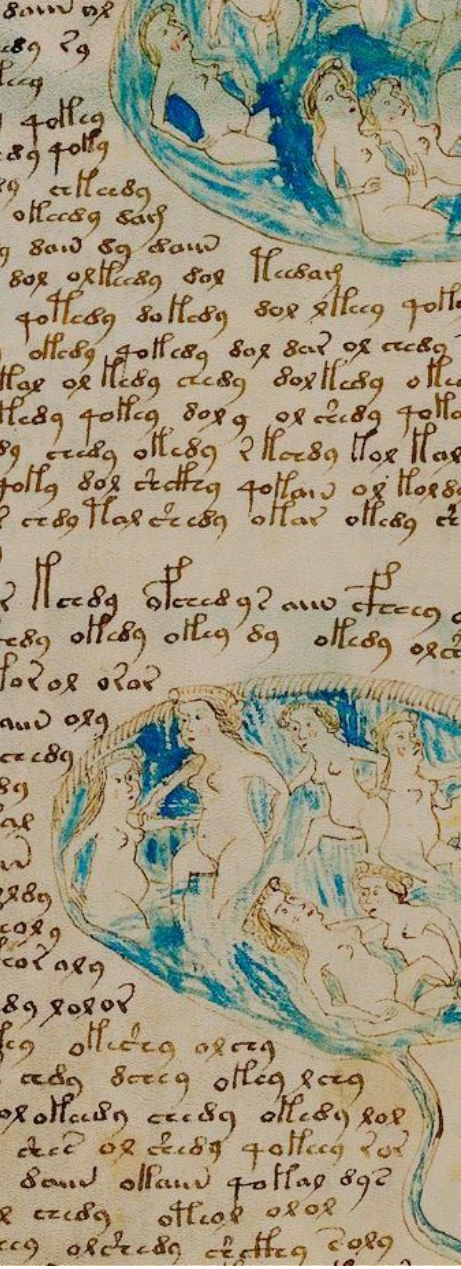
- Unknown Origin
- Unknown Author
- Unknown Writing System
- Unknow Language or Language Family

WHY IS IT INTERESTING?

The manuscript has been systematically studied by professional and amateur linguists and cryptographers alike

Many have made decipherment claims but the book has never demonstrably decoded

It is still entirely unclear if the book is a Hoax, is written in a plain Natural language, or is Enciphered!



“DECODED”

Hebrew

Voynich Manuscript SOLVED: World's most mysterious book deciphered after 600 years

Latin

Has the Voynich Manuscript Finally Been Decoded?: Researchers Claim That the Mysterious Text Was Written in Phonetic Old Turkish

in Books, History | February 21st, 2019 19 Comments

Turkish

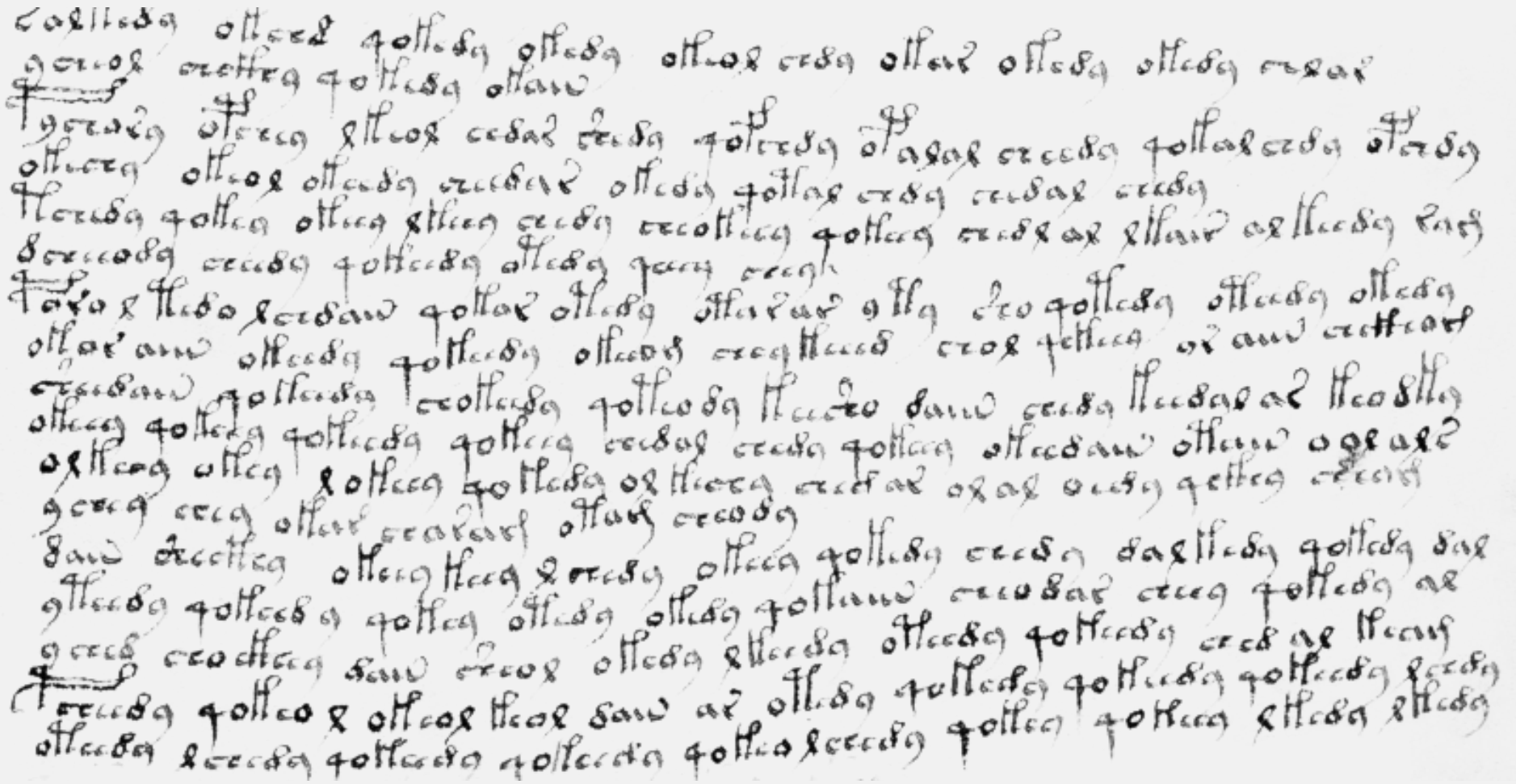
21 FEBRUARY, 2014 - 21:30 APRILHOLLOWAY

First words in mysterious Voynich Manuscript decoded

Arabic

Linguists often use “circular” and self-reaffirming thought processes to try to translate words in the manuscript

HOAX, CODE, OR NATURAL LANGUAGE?



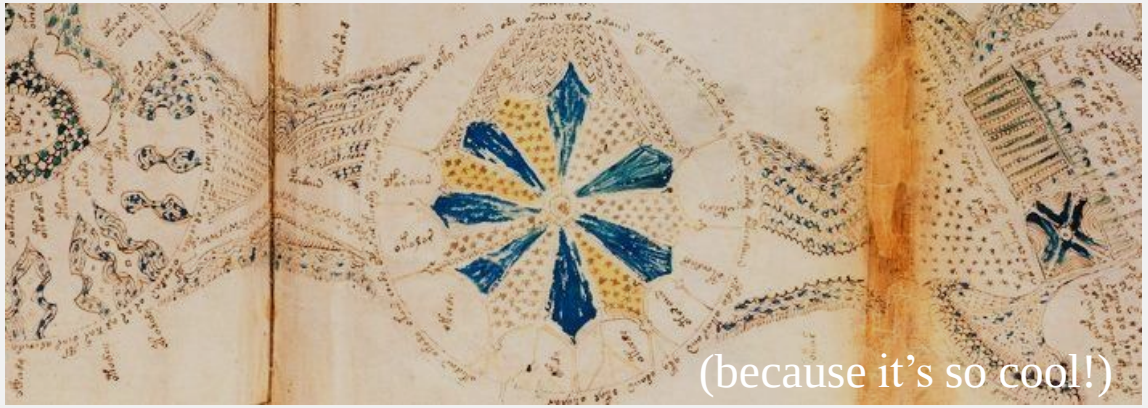
- What tools to use?
- How can we identify a natural language?

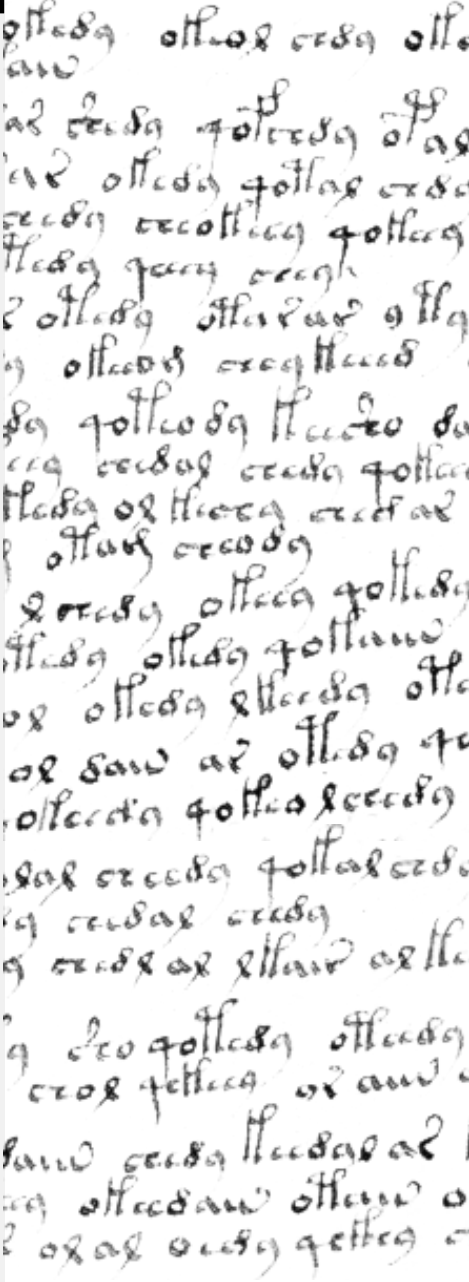
HOAX, CODE, OR NATURAL LANGUAGE?

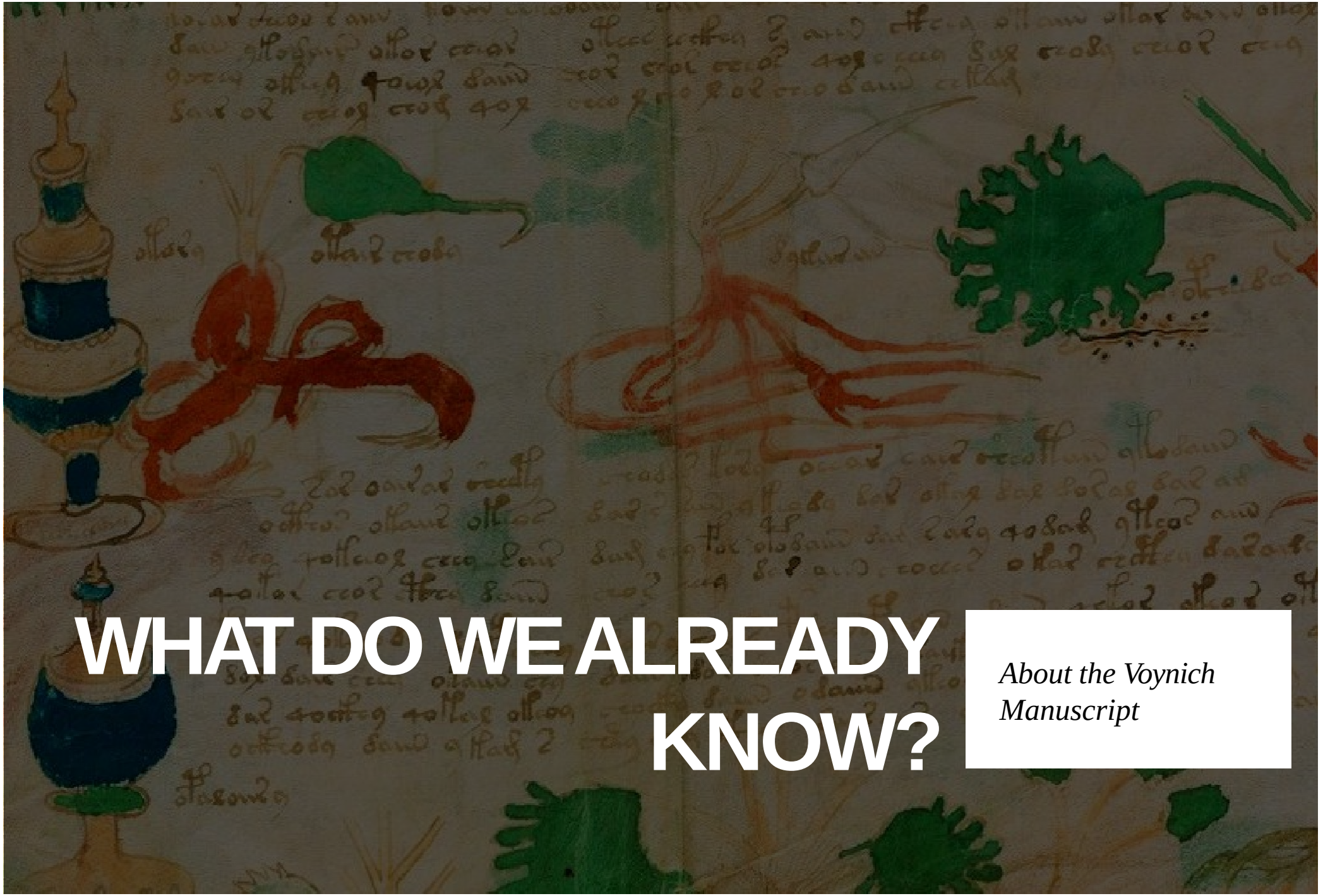
- What tools to use?
- How can we identify a natural language?

New options might be open now with advances in NLP

My main point is to share my big interest for this book with other NLP researchers so that more ideas might surge for exploring the properties of this book







WHAT DO WE ALREADY KNOW?

About the Voynich Manuscript

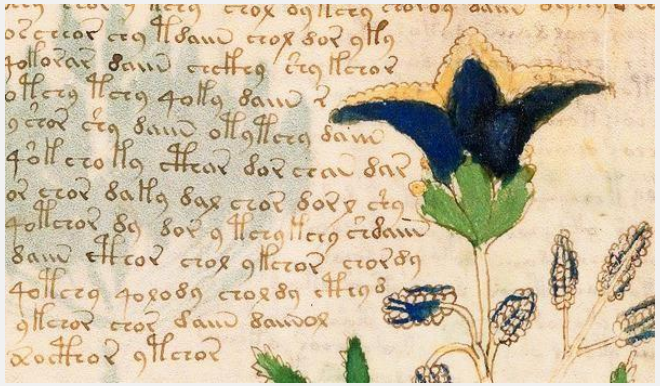
ORIGINS OF THE MANUSCRIPT

- The vellum was carbon dated to 1404-1438
- Made with techniques that where common all over Europe
- Uses pigments that where cheap and common in Europe
- Earliest letter about its existence suggests it was sold to Rudolph II, emperor of the Holy Roman Empire, in the late 16th century
- Nothing is known about the almost 200 years between its creation and sale
- Sold to him for 600 ducats (pure gold coins) (about 70 thousand euro worth of gold today)

CONTENTS OF THE MANUSCRIPT

Herbal

(drawings of plants)

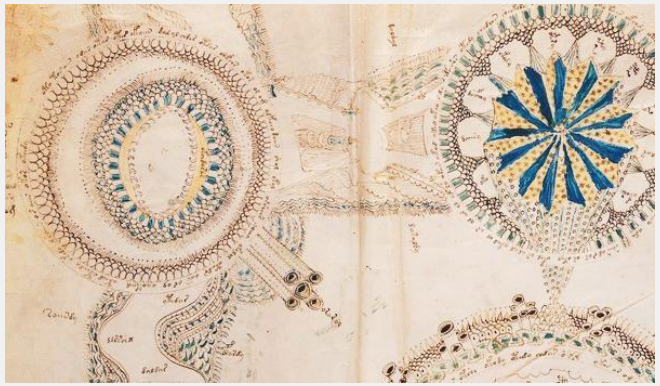


Astrological

(zodiac star maps)

Biological

(bathing women)

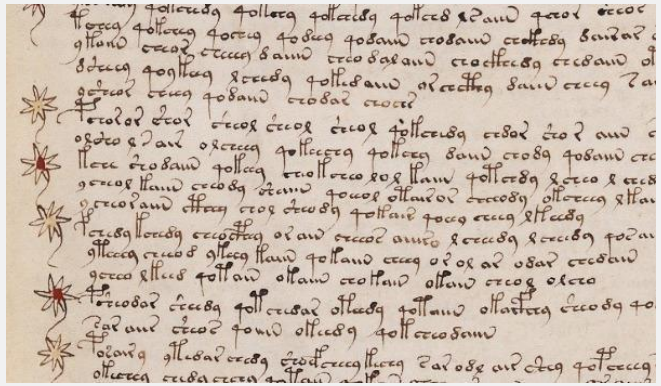


Cosmological

(circular illustrations)

Pharmaceutical

(parts of plants)



Recipes

(itemized text)

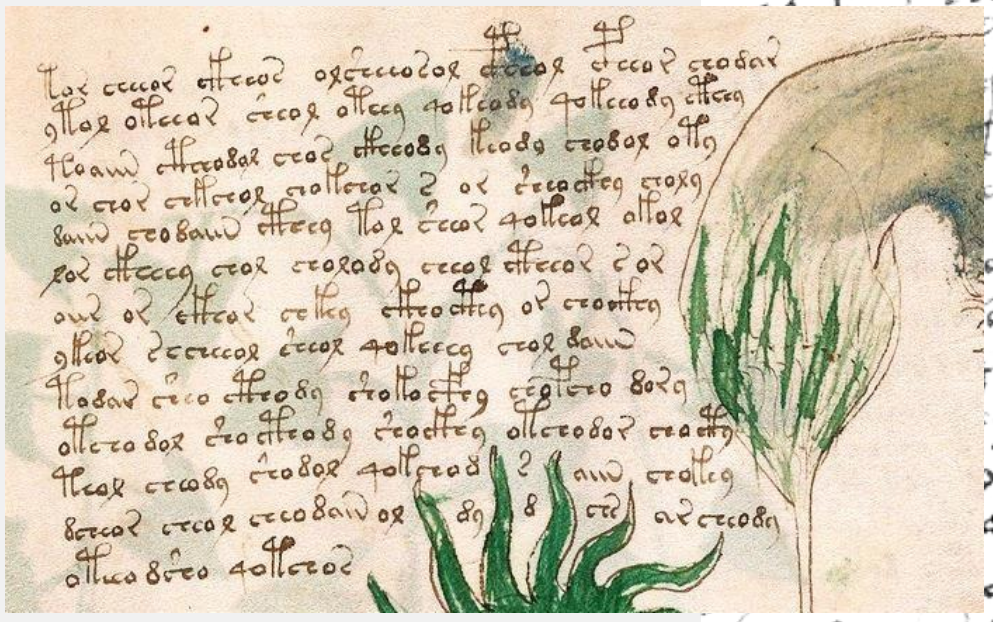
Based on illustrations, the book can be divided into 6 different sections

THE TEXT

The language is often called “Voynichese” and is visually similar to many European scripts

Transcriptions of the text use different numbers of letters (depends on how some “connected” symbols are interpreted)

About 35 common and frequent letters with as many as 200 rare letters



a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

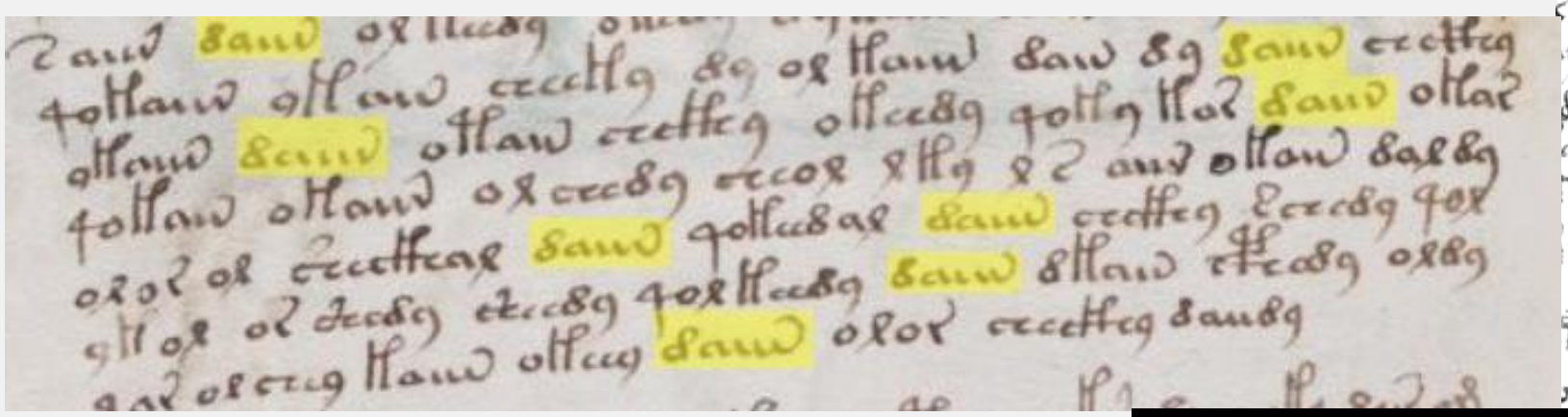
OTHER QUICK OBSERVATIONS

- There is not a single instance of a word being erased or corrected (which suggests the book might be a copy) but some words have been retouched
- Serious cryptography systems (that where not simple letter reordering or substitution ciphers) did not reach Europe until years after the manuscript was written
- Some pages appear to have different handwritings. As many as 6 scribes are said to have been involved in writing the book

STATISTICAL PROPERTIES

Of the Voynich Manuscript

WORD TOKENS



Sample text with most common word in the book labelled

170 000 characters (approximately)

37 919 word tokens

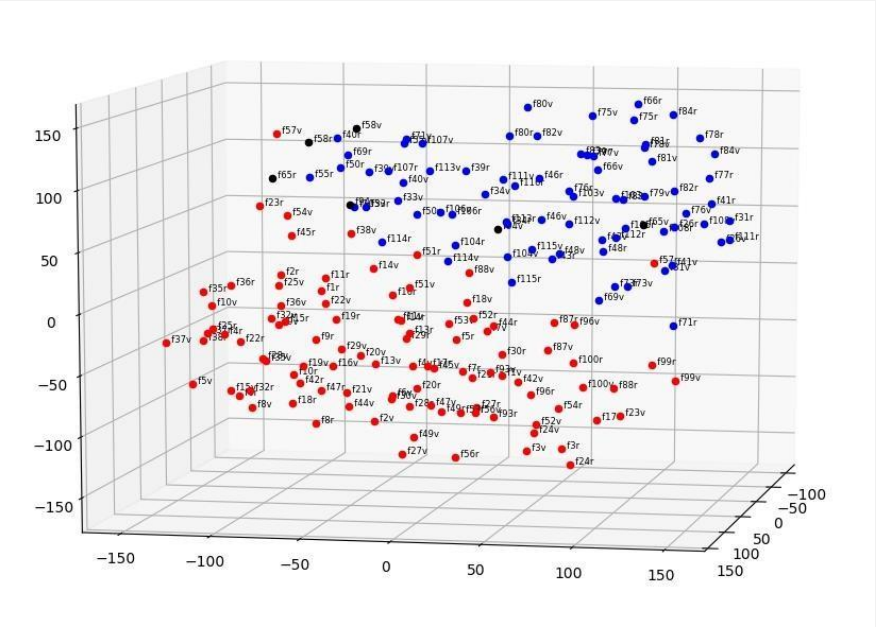
8 114 unique word tokens

POSSIBLE DIFFERENT LANGUAGES

One of the first things that where first noticed about the first and second order statistics of the text was that they where sometimes very different between two pages

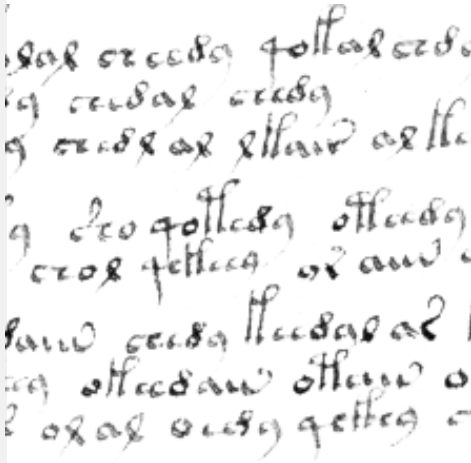
As indicated by P. Currier, the pages are written in one of two languages (Currier A, or Currier B) with distinct statistical properties and rules for word formation

The two languages have also been associated with different handwritings



The two languages might also be different dialects, different ciphers, language domains

For NLP purposes this means that our already small dataset of Voynichese is actually two smaller datasets!



PROBABLY DIFFERENT LANGUAGES

Words common in Language A, but rare in Language B

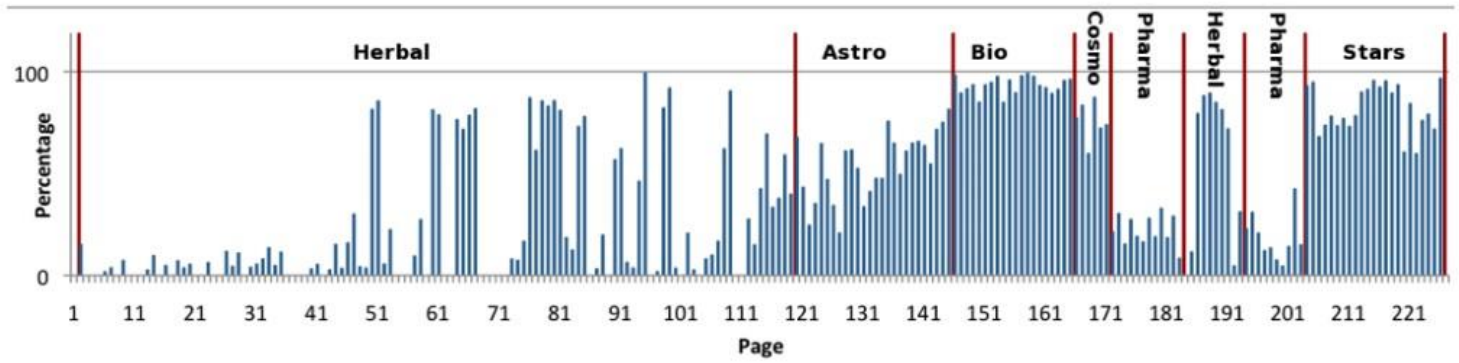
Word	# in Language A	# in Language B
	61	1
	20	1
	14	1
	28	2
	13	1
	13	1
	13	1
	13	1
	11	1
	21	2

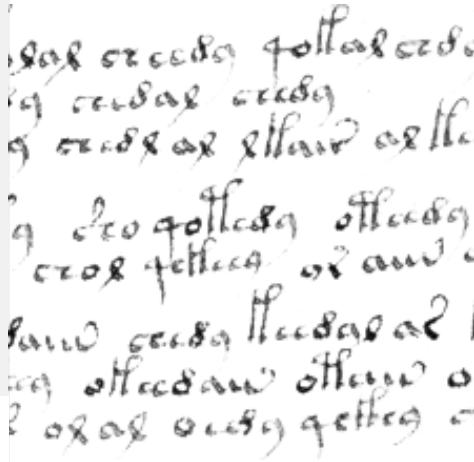
Words common in Language B, but rare in Language A

Word	# in Language B	# in Language A
	159	1
	112	1
	107	1
	50	1
	28	1
	21	1
	81	4
	40	2
	18	1
	31	2

Most research focuses on Currier B due to the larger number of pages

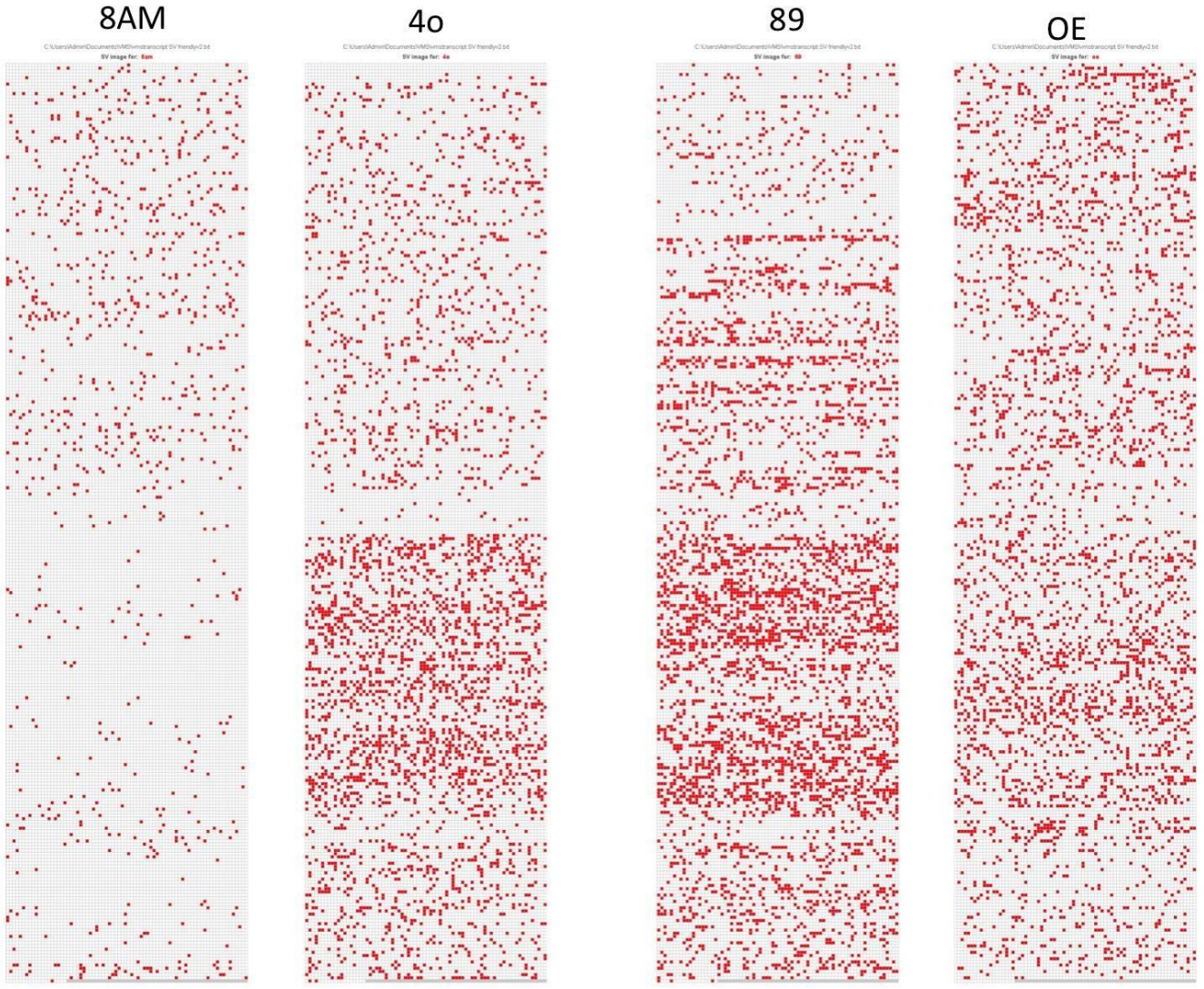
Figure 2: VMS sections, and percentage of word tokens in each page that are tagged as language B by the HMM.





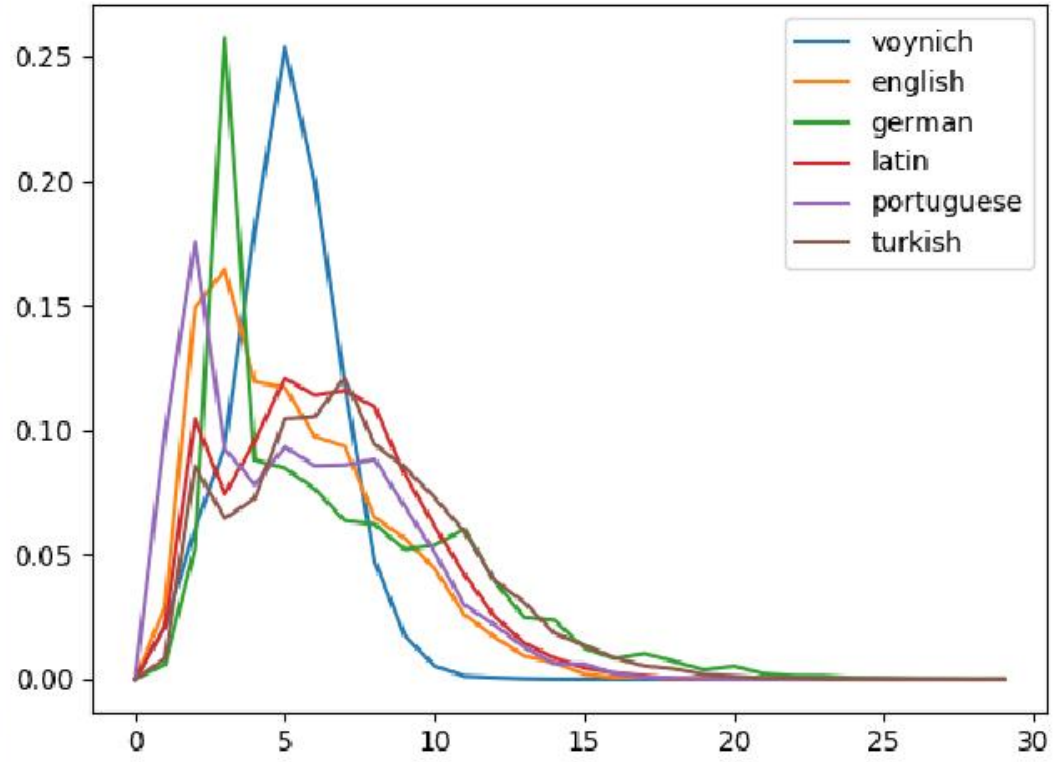
PROBABLY DIFFERENT LANGUAGES

Common syllables in the Voynich Manuscript, showing entire manuscript

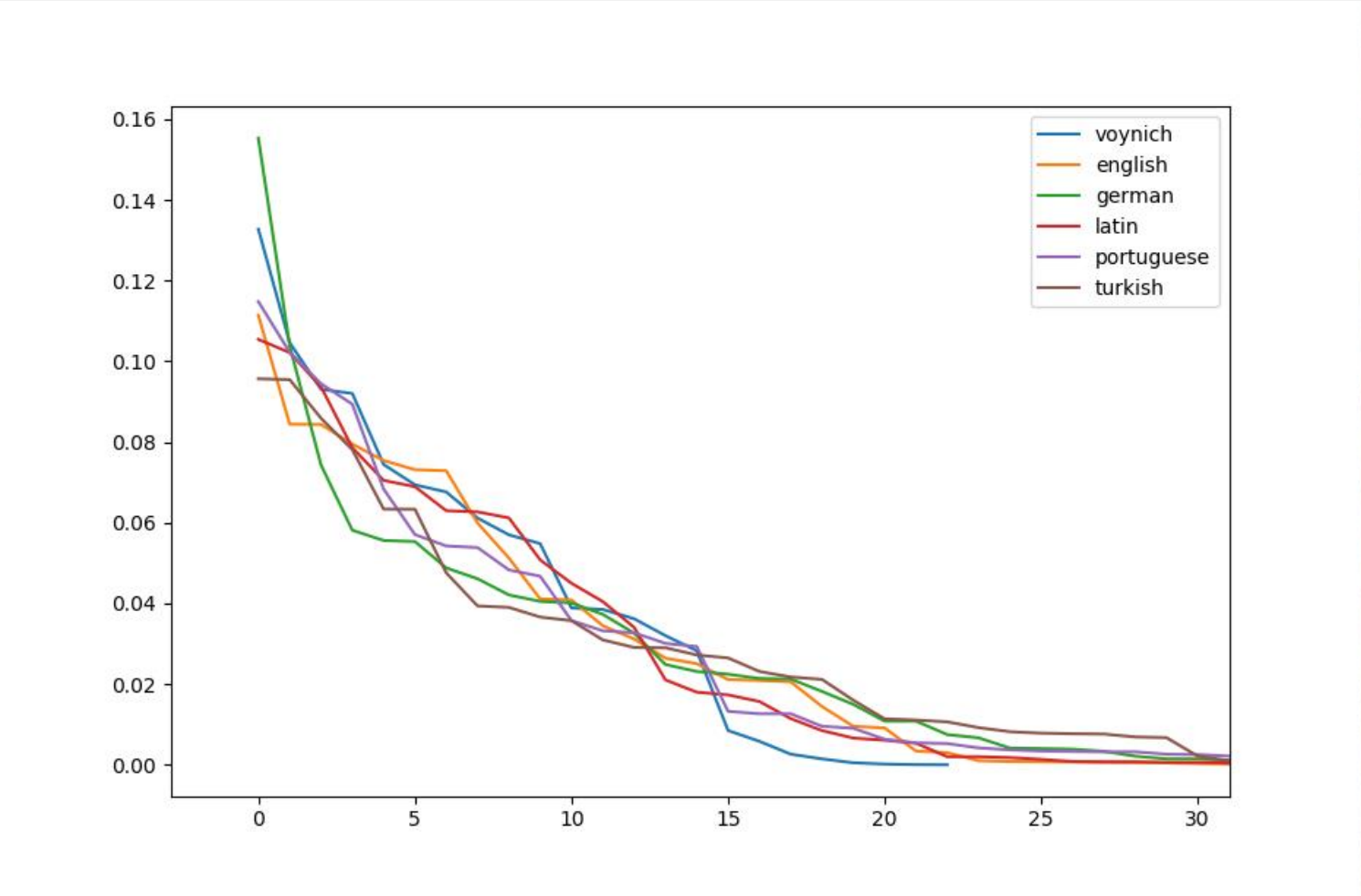


Certain syllables are very frequent in only one of the languages

WORD SIZES SEEM VERY NATURAL

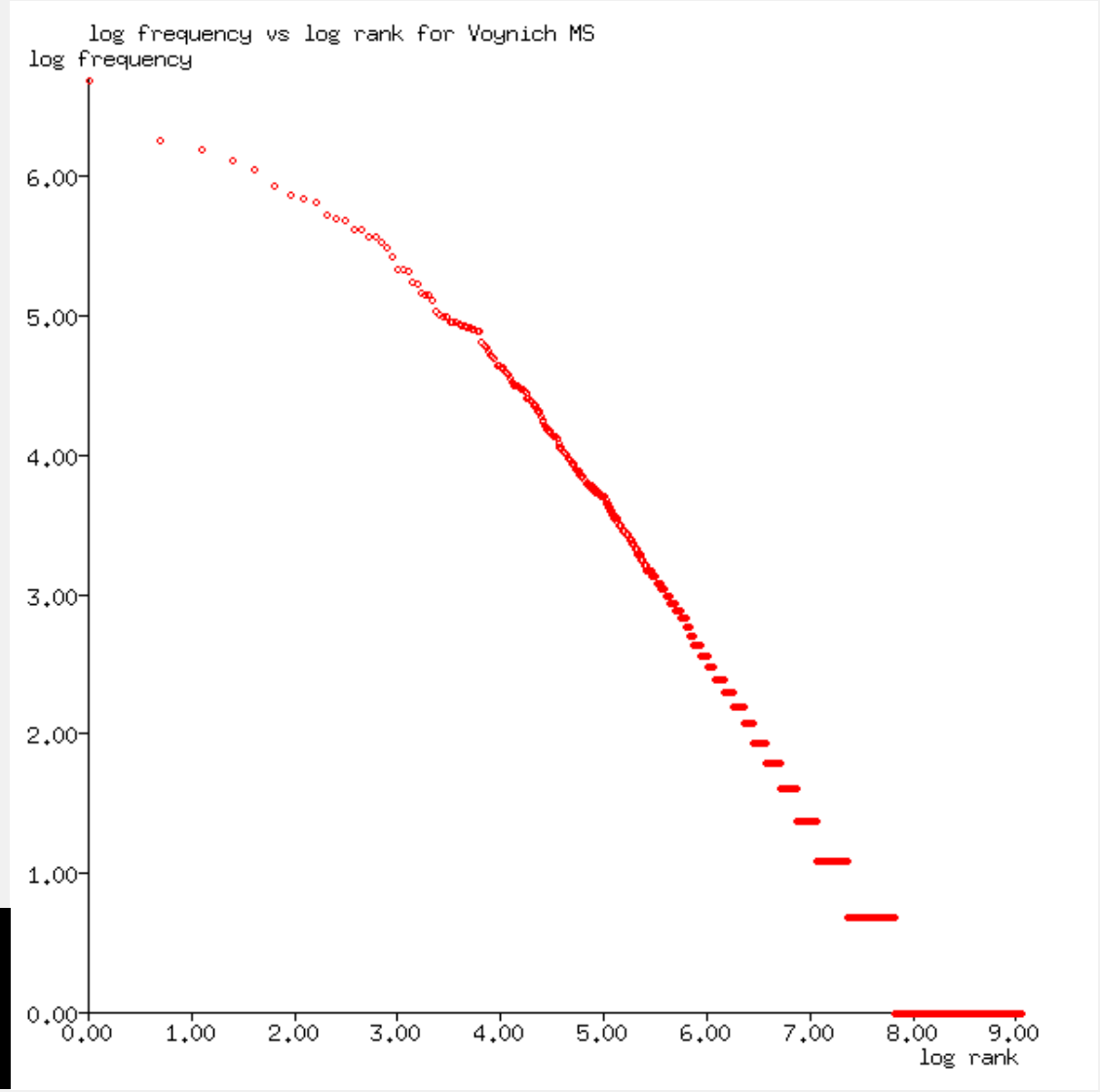


LETTER FREQUENCIES ARE ALSO NATURAL



WORD FREQUENCIES FOLLOW ZIPF'S LAW

Zipf's law for word length is also followed by Voynichese (shorter words are more frequent)



This is expected of a real natural language

ARE THERE VOWELS AND CONSONANTS?

Sukhotin's algorithm for vowel identification (works perfectly on modern languages) separates 4 symbols as vowels but with a low degree of separation, and many words do not contain any of these symbols

Expectation Maximization algorithms for a two-state Hidden Markov Model yields similar results

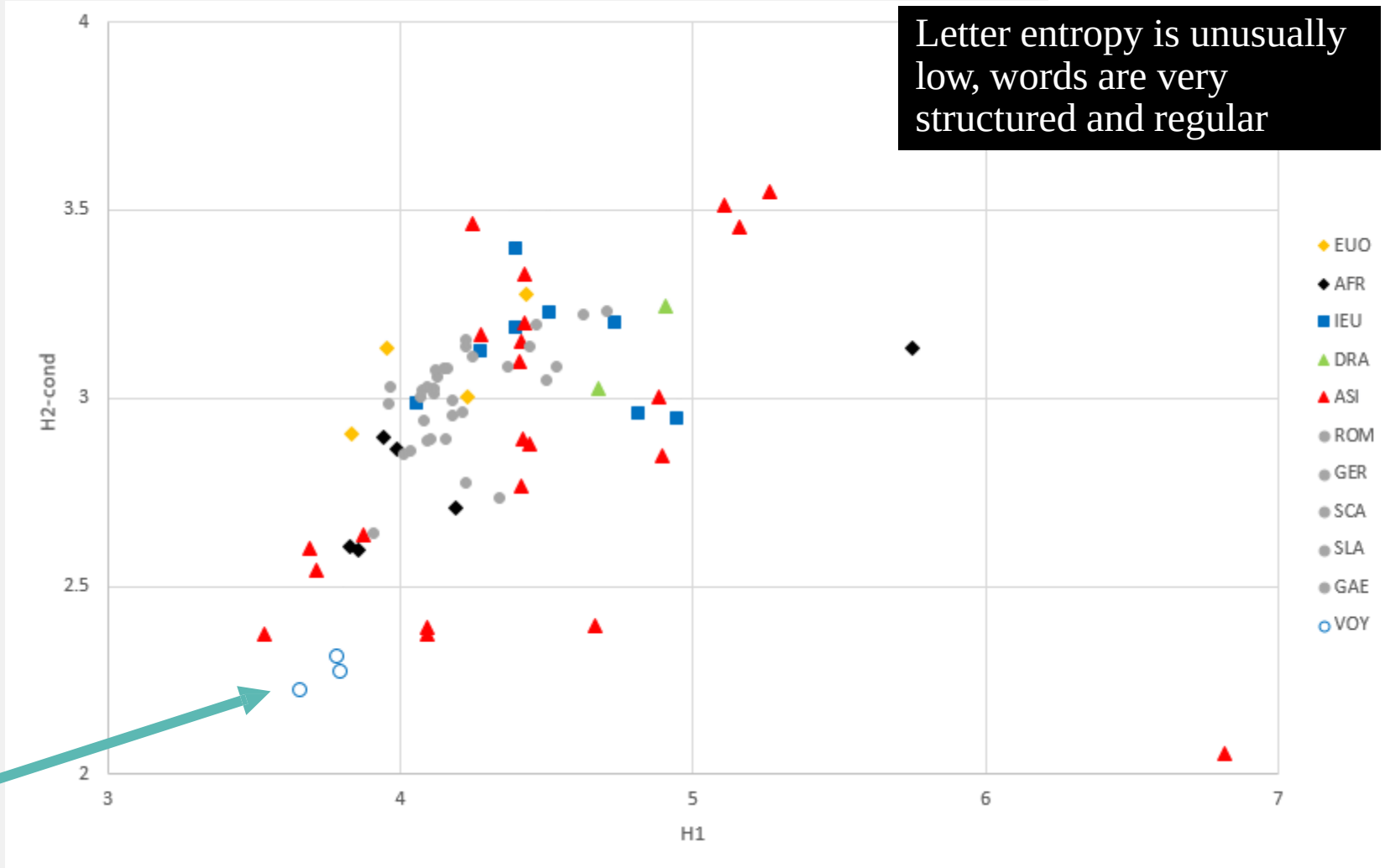
Further research has shown that the letters can be grouped into two classes: letters that appear at the end of words, and all others

This is a weird property not shared with modern languages

RegEx grammar for word construction is basically a*b

WORD STRUCTURE AND LETTER ENTROPY

Code	Meaning
ROM	Romance Languages
GER	Germaing Languages
SCA	Nordic Languages
SLA	Slavic Languages
GAE	Gaelic Languages
EUO	Other European
AFR	African
IEU	Other Indo-European
DRA	Dravidian
ASI	Asian
VOY	Voynichese



Letter entropy is unusually low, words are very structured and regular

GRAMMATICAL STRUCTURE AND WORD ENTROPY

The text has famously low quantities of repeated word pairs and trios

The text feels very random in word order, the current word shows little to no information about the next

Table 4: Predictability of words (over 10-fold cross-validation) with bigram contexts, compared to unigrams.

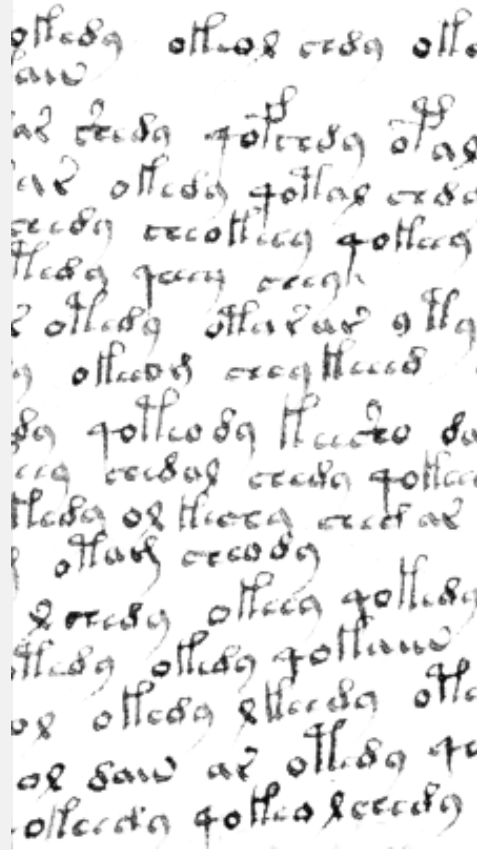
	Unigram	Bigram	Improvement
VMS B	2.30%	2.50%	8.85%
English	4.72%	11.9%	151%
Arabic	3.81%	14.2%	252%
Chinese	16.5%	19.8%	19.7%
Hungarian	5.84%	13.0%	123%

Less Predictability

=

More Entropy

None of the tested natural languages had such a high word bigram entropy



PAGE TOPICALITY

If pages have specific topics, it is expected that there is an unusually higher frequency of some words

Table 6: Strength of page topics in VMS and other texts, cropped to be of comparable length to the VMS.

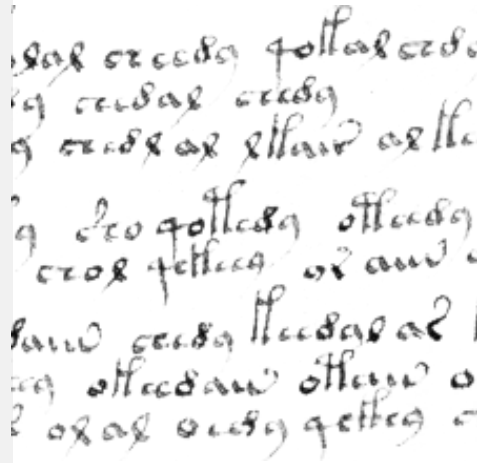
	VMS B	English WSJ	English Genesis	Arabic Quran
T	7.5	6.3	6.6	7.7
T_{rand}	7.7	6.5	7.1	7.9
$1 - T/T_{rand}$	0.033	0.037	0.069	0.025

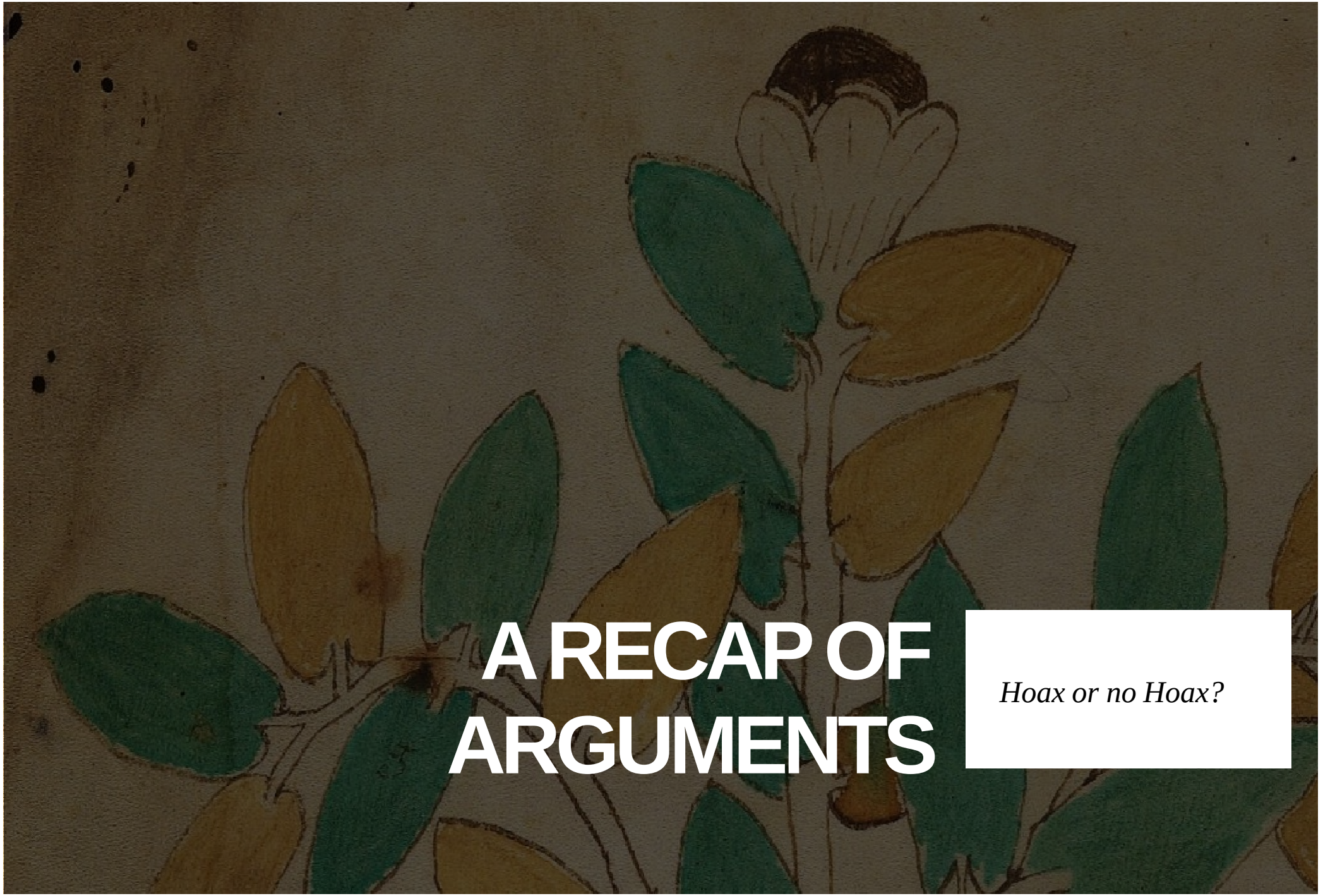
For reference, the topicality of pages (calculated with word entropy) is compared to the topicality of the scrambled document

Adjacent pages share a lot of common vocabulary, as expected of a real book with ordered pages (calculated here with bag of words cosine similarity)

Table 7: ADJPAGESIM for VMS and other texts.

VMS B	38.8%
VMS All	15.6%
VMS B pages scrambled	0%
VMS All pages scrambled	0.444%
WSJ	1.34%
English Genesis	25.0%
Arabic Quran	27.5%

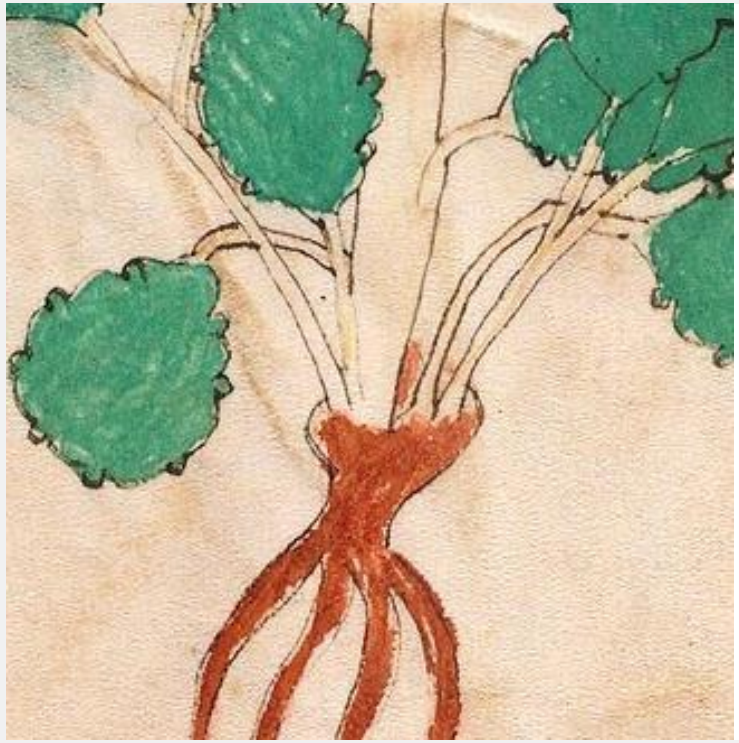




A RECAP OF ARGUMENTS

Hoax or no Hoax?

WHY IT'S PROBABLY A HOAX



- No one has ever been able to decipher it
- Almost completely random word order with no grammatical structure
- The lack of any instance of a word being corrected
- Sold in ancient times for a lot of money which would justify creating it

WHY IT'S PROBABLY NOT A HOAX

- Follows Zipf's law for word frequencies
- Natural word sizes and letter frequencies
- The two identified languages/dialects
- Pages have topics, nearby pages have similar topics
- A book of this size would be very hard to fill with fake text

LEADING HYPOTHESIES

For origin

IF NOT A HOAX

Cypher: Could be an existing European language intentionally masked behind some cypher

Shorthand: Could be stenography for efficient writing of an existing language. Individual symbols could replace long words

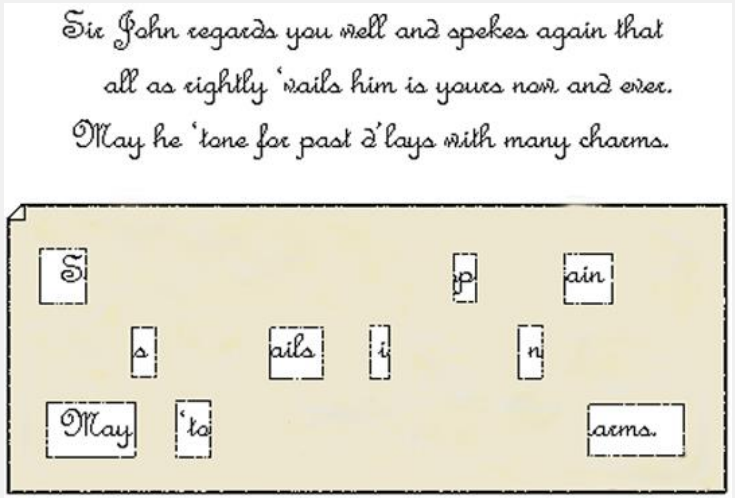
Lost Natural Language: Could be the last written evidence of some minor language that has gone extinct

Phonetic Writing: Could be a known or lost language written mostly phonetically due to lack of a consistent writing backlog. Different people might write the same words as different sounds (which would explain the two apparent languages)

IF A HOAX

Why: The book was sold for a lot of money

How: Either by trying to write random letters, or by using dice or some other system for quickly creating random words, or by trying to shuffle a real text in a real language



A Cardan grille, used for hiding messages, could have had been used to create the text

WHAT TO DO?

As Computational Linguists

RESOURCES

For Study and Research

See the book (includes mapped transliteration and word finder):

http://www.voynichese.com

Text descriptions of the images and comparison of all known transliterations:

http://voynich.freie-literatur.de/index.php

Most popular site for Voynich resources, includes download links for transcriptions:

http://www.voynich.nu

Article summarizing some of the known properties of the book:

https://www.isi.edu/natural-language/people/voynich-11.pdf

Github repository I made with one of the transcriptions and a python parser for it:

https://github.com/joaoperfig/voynichstudies


THANK YOU!