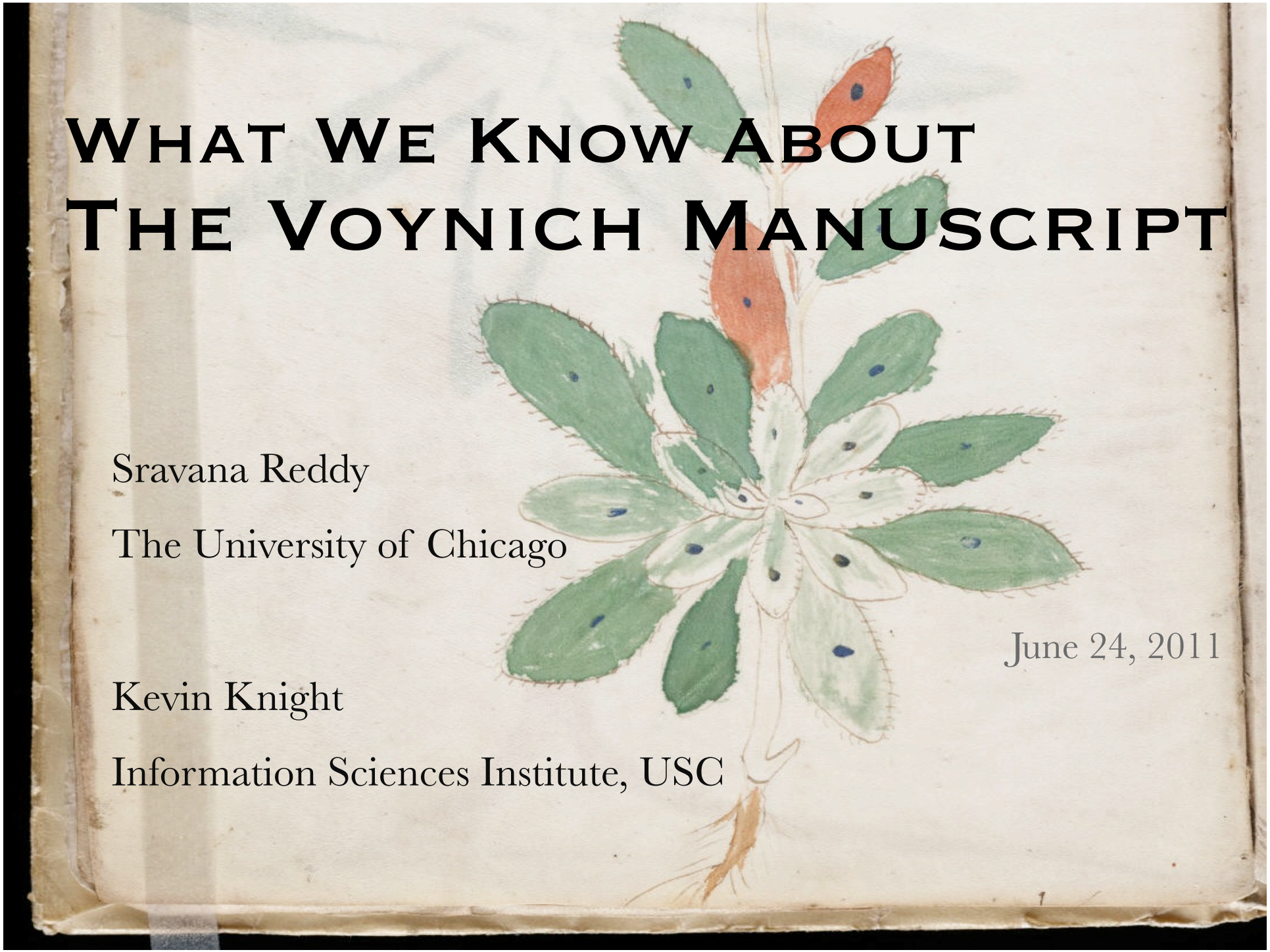


What We Know About The Voynich Manuscript

Sravana Reddy

The University of Chicago

June 24, 2011

Kevin Knight

Information Sciences Institute, USC

Background

- Medieval manuscript; undeciphered script
- ~235 pages, 38000 words
- 25-40 characters
  ( = single char, or + ?, etc)
- Based on illustrations, divided into 5 sections:
  - *Herbal*

Background

- Medieval manuscript; undeciphered script
- ~235 pages, 38000 words
- 25-40 characters
  ( = single char, or + ?, etc)
- Based on illustrations, divided into 5 sections:
  - *Herbal*
  - *Astrological*

Background

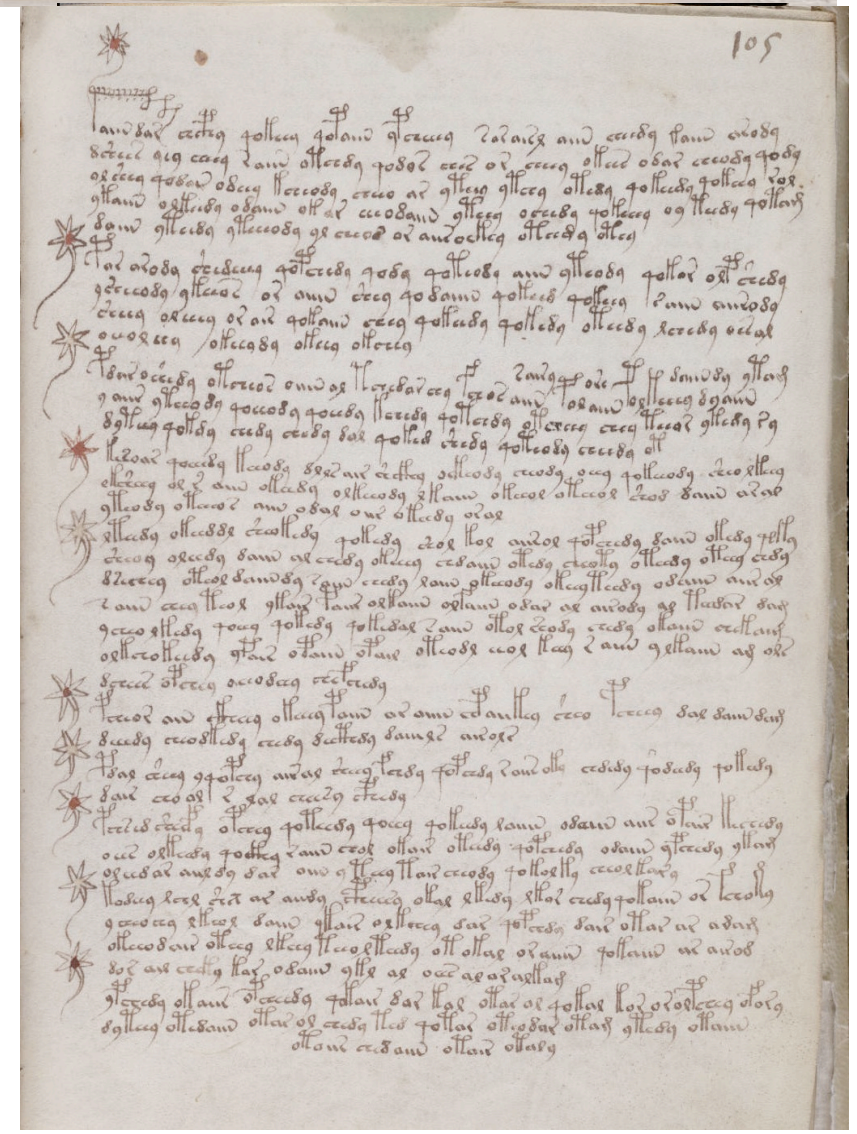
- Medieval manuscript; undeciphered script
- ~235 pages, 38000 words
- 25-40 characters
  ( = single char, or  +  ?, etc)
- Based on illustrations, divided into 5 sections:
  - *Herbal*
  - *Astrological*
  - *Biological*

Background

- Medieval manuscript; undeciphered script
- ~235 pages, 38000 words
- 25-40 characters
  ( = single char, or +  ?, etc)
- Based on illustrations, divided into 5 sections:
  - *Herbal*
  - *Astrological*
  - *Biological*
  - *Pharmacological*

Background

- Medieval manuscript; undeciphered script
- ~235 pages, 38000 words
- 25-40 characters
  ( = single char, or +  ?, etc)
- Based on illustrations, divided into 5 sections:
  - *Herbal*
  - *Astrological*
  - *Biological*
  - *Pharmacological*
  - *Stars*



History

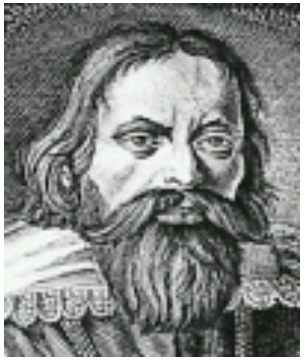
(Roger Bacon)

John Dee (England), 16th century

Emperor Rudolph II (Austria), 16th century

History

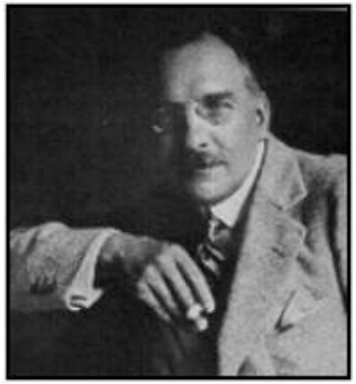
Jacobi de Tepenecz
(Prague),
16th/17th century



Jan Marek Marci
(Prague),
17th century

Athanasius Kircher
(Germany),
17th century

History



Wilfrid Voynich
(Italy), 20th century

H. P. Kraus
(USA)

Yale Library

Carbon-dating of paper at University of Arizona (unpublished): 15th century
Dating of ink at McCrone Research Institute: added soon after the paper

Motivation

- Decipherment of the manuscript could make you famous
- (Computational) Linguists have the right tools + understanding of text analysis – so we should be the ones looking at it
- No gold standard or answer key.
  Good opportunity to try out unsupervised algorithms.

Questions

Is there syntax? Word order?

Part-of-speech categories?

Long-distance collocations?

Are there vowels and consonants?

Do letters have cases?

How many authors?

Is there punctuation?

What are the word frequency and length distributions?

Is there morphology?

How predictable are letters within words?

Is there a narrative? Topics?

Is it prose?

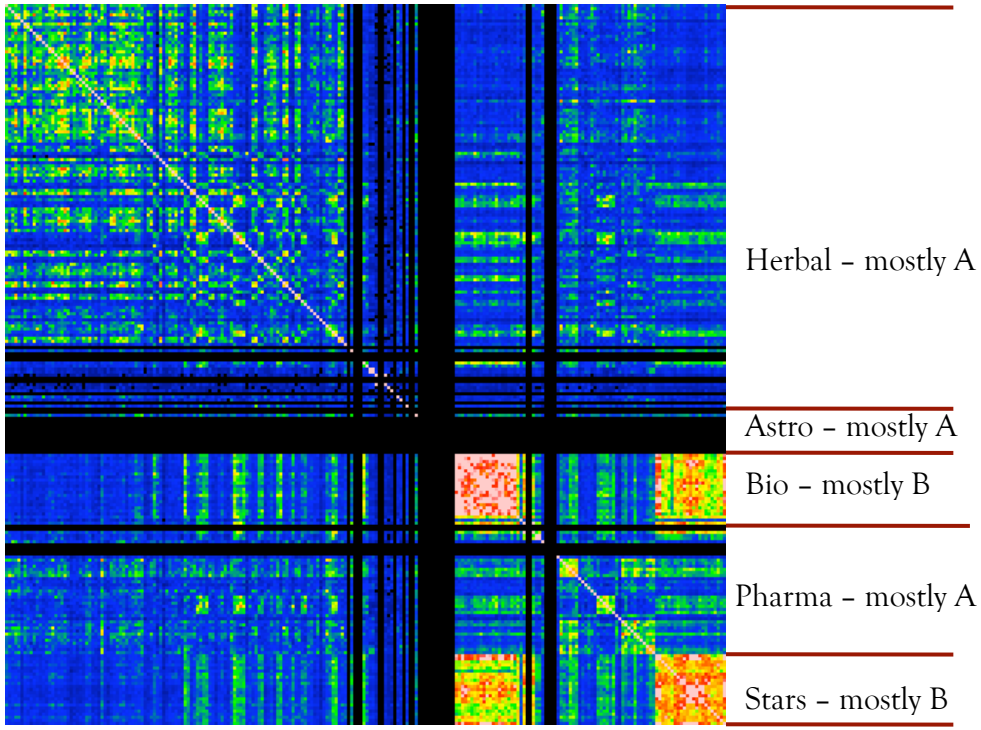
Does the text correlate with the illustrations?

Are the pages in order?

Is it a language, a code, or a hoax?

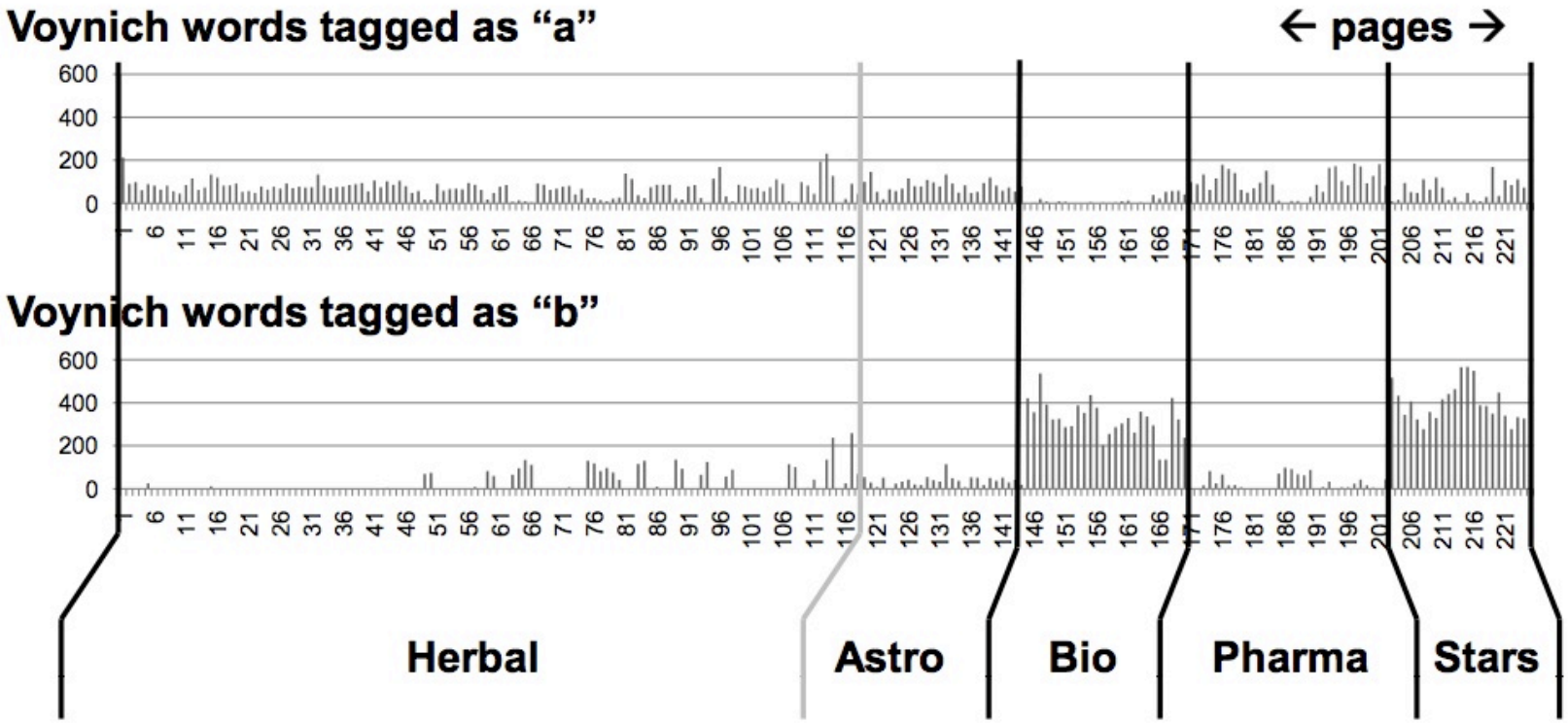
How Many Authors?

- Two distinct 'languages', A and B
  - Currier (1976) – vocabulary and handwriting differences
  - Zandbergen (1997) – page-similarity plot



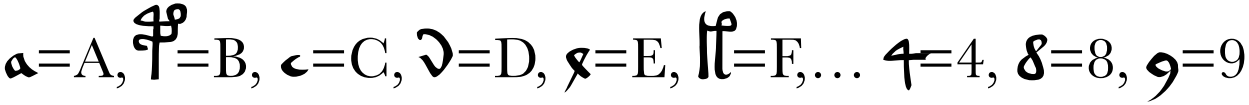
How Many Authors?

- Two distinct 'languages', A and B
  - We clustered words into 2 classes with bigram HMM



Data

- Currier/D'Imperio transcription – 35 characters

- For coherence, most of analysis is on "Voynich B" – sections written in the B language (Bio and Stars)
  - 19415 word tokens, 49 pages

- Texts for comparison:
  - Part of the Wall Street Journal (WSJ) corpus (28551 words)
  - Part of Arabic Quran, no diacritics (19327 words)
  - Chinese Sinica Treebank (18791 words)

Are There Vowels and Consonants?

- If so, we can find out which characters are consonants and which are vowels (Sukhotin, 1962; Knight et al, 2006; Goldsmith & Xanthos; 2009)
- Intuition: All vowels occur in the same general contexts (and similarly for consonants)
- Using EM with two states, we found:

a b _ b b a _ b a b b _
i n _ t h e _ t o w n _

b b a b a _ a _ …
w h e r e _ i _ …

b b b b a _ a b b a _ b a _
V A S 9 2 _ 9 F A E _ A R _

b b b a _ b b a _ b b b a _ …
A P A M _ Z O E _ Z O R 9 _ …

Are There Vowels and Consonants?

- Possible explanations:
  - Last character is vowel – placed at end of word
  - Characters are syllables or morphemes
  - Abjad like Semitic scripts – most vowels not written

But even long words seem to have only one 'vowel'!

Only 35 characters...

```
a a _ b b b a _
t h _ D t c h _

b b b a a a a _ b b b a
p b l s h n g _ g r p .
```

Devoweled English

Most likely!

```
b b a _ b b b a _
b s m _ A l l h _

b b b b a a _ b b b b b a
A l r h m n _ A l r h y m
```

Arabic with no diacritics

Do Letters have Cases?

- Some characters only occur at beginnings of paragraphs/lines

  𐤉𐤌𐤌𐤃

- Decorative uppercase?

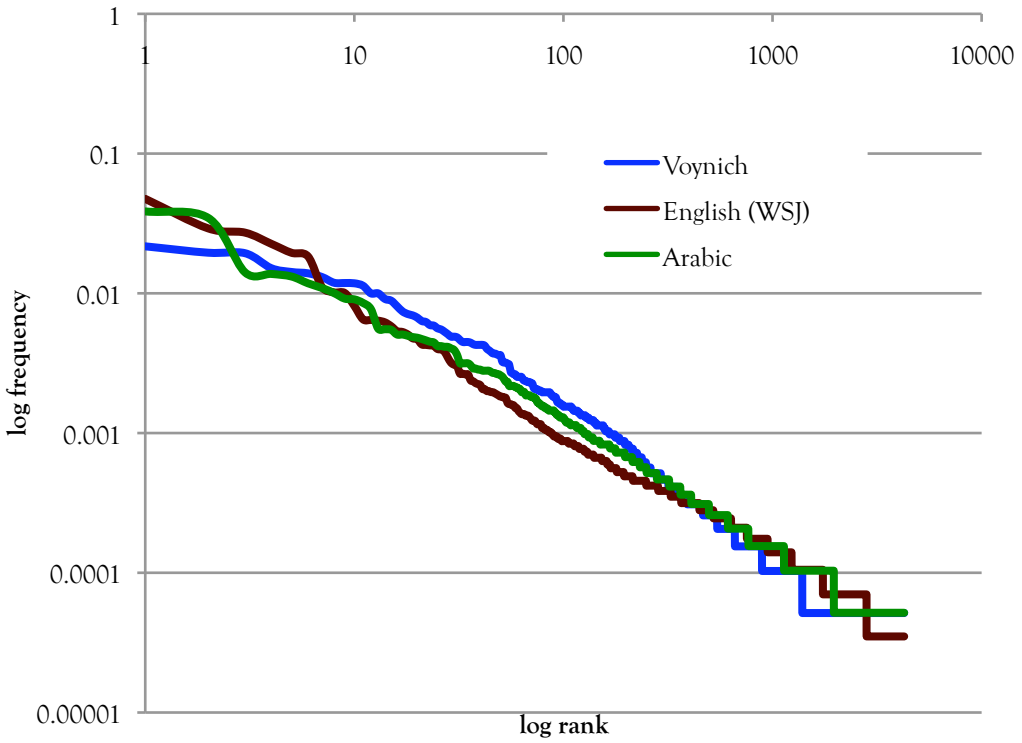
- To find lowercase equivalent of A in English:
  - replace every instance of A with another character
  - compute decrease in word entropy
  - character with highest decrease is lowercase equivalent

- But lowercase of 𐤉 is 𐤃, 𐤌 is 𐤃, etc.

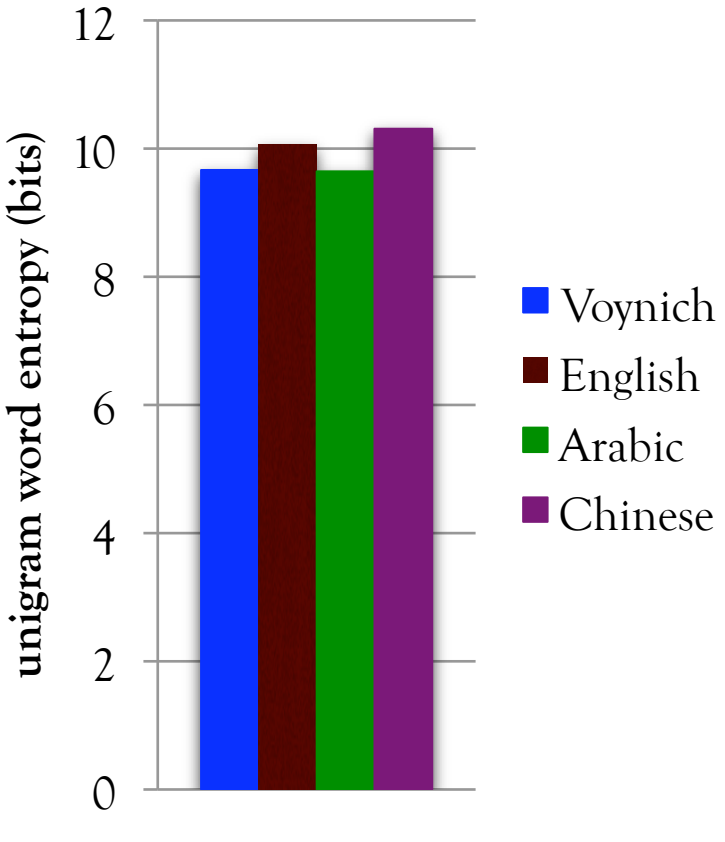
Is there Punctuation?

- Definition:
  - Characters that occur exclusively at ends or beginnings of words
  - Removing character results in a word
- No such characters in manuscript

Word Frequencies?

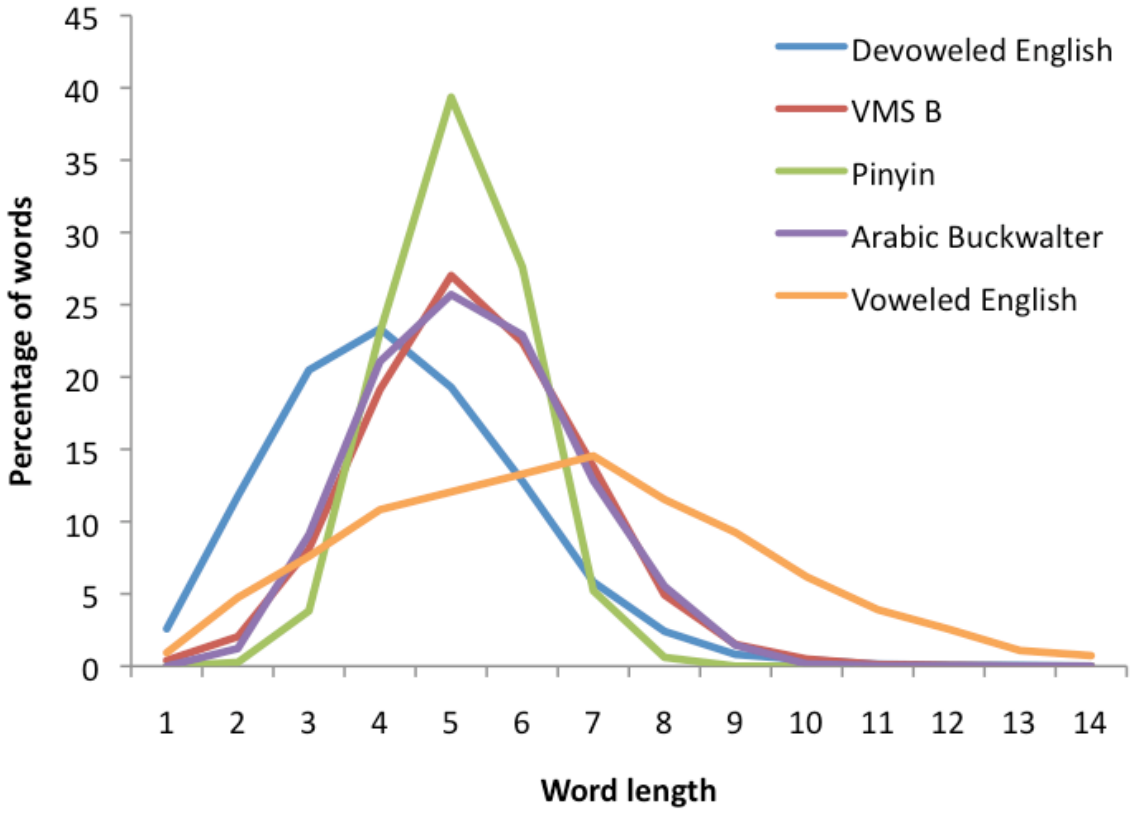


Frequency/rank distribution



Unigram word entropy

Word Lengths?



- Binomial distribution has been observed, and perceived as unnatural.
- But clearly, distribution is reasonable for no vowels.

Predictability of Letters

- Bigram predictability: how well can you guess a character given the previous one?
- Average % accuracy over 10-fold cross-validation:

	Voynich	English	Arabic	Pinyin
Bigram predictability	40.02	22.62	24.78	38.92
Unigram predictability	14.65	11.09	13.29	11.20

Is there Morphology?

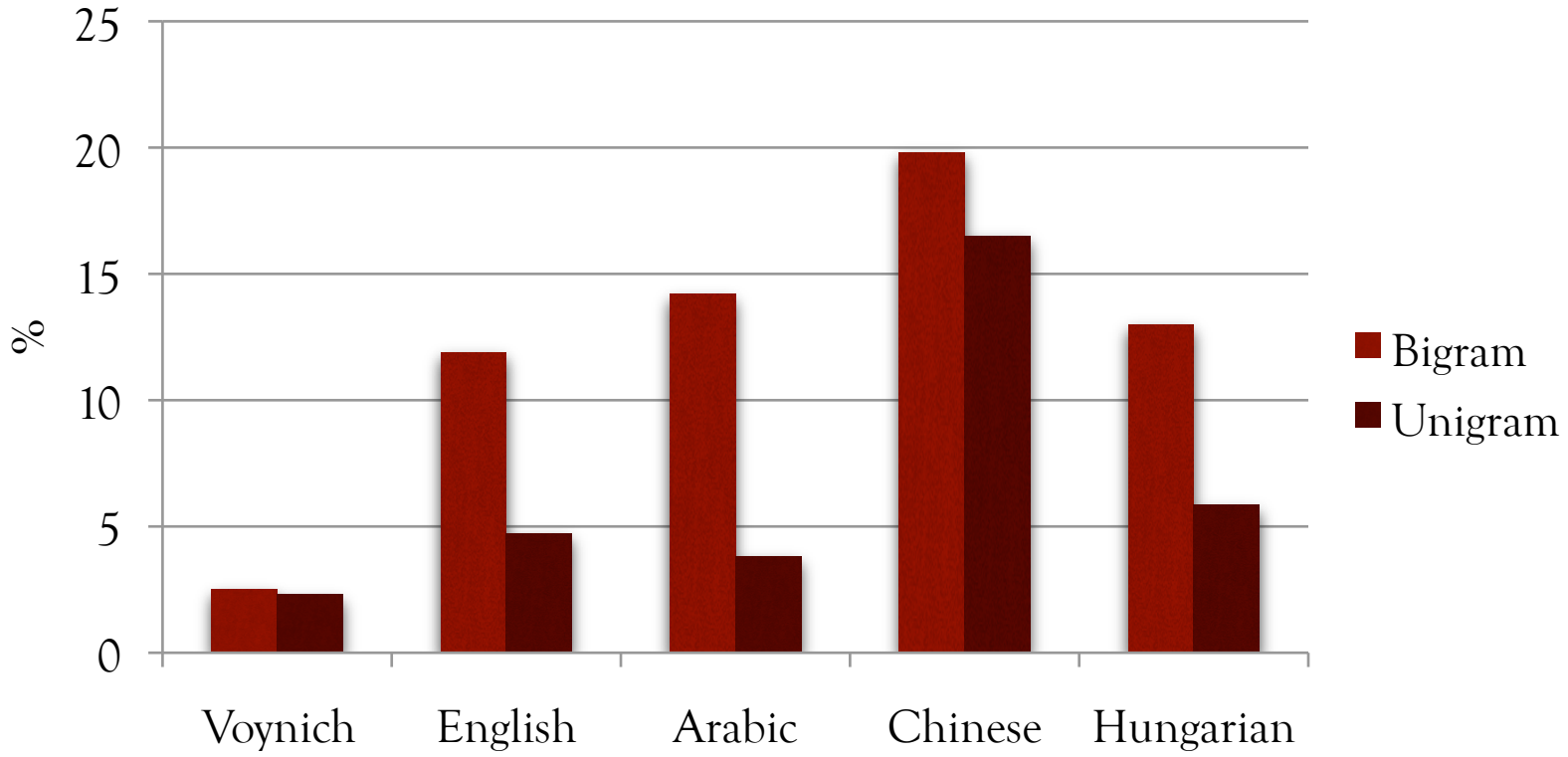
- Unsupervised morphological analyzer Linguistica (Goldsmith 2001): MDL to find prefix+stem+suffix segmentaion
  - 'Signatures': groups of affixes that take the same set of stems

Affixes	Stems
OP+	CAE CAM CAN CC2AE CC6 CC8 CC8AE CC8AN CC8AR CC8SC9 CC9FZ9 CCAE CCAJ CCC8SC9 CCC9 CCCO2 CCO2 CCO8AG CCO8AJ CCO8E CCOEFCC9 CCOESOR CCR CCS9 OE89E OEOP9...
+89, null	OFAJ 4OFC89 4OFCAE 4OFCCO 4OFCO 4OFO 4OFOE 4OPAE 4OPAR 4OPCC8 4OPCCO 4OPCO 4OPS2 4OPSO 8AE 8AM 8AT9 8SCO 8ZCO 9FCCO 9PCCO 9SCCO 9SCO 9ZCAE 9ZCCO EFCCO EFCO EFCOE EFE EO2 EOE EZCO FAE FCCCO FCCO FCCZO FOE...
OE+, OP+, null	8AE A3 AD AE AE9 AEOR AJ AM AN AR AT E O O2 OE OJ OM ON OR SAJ SAR SCC9 SCCO SCO2 SO
+9, +C89	4CF 4CP 4OS 4X 89P 89PZ 89S 8AEF 8ARS 9FCCZ 9P AEZ AFS EF ESCX EZCX FOEF OCFC OEBZ OEF OFCZ OW OX PSOF Q ROES ROEZC RS SBS SC9F SCBS SCFC SCOQ SOFS SOQ SOX SP W X ZCBS ZCOFC ZCQ ZFC ZFS ZX...

Stems have similar spellings

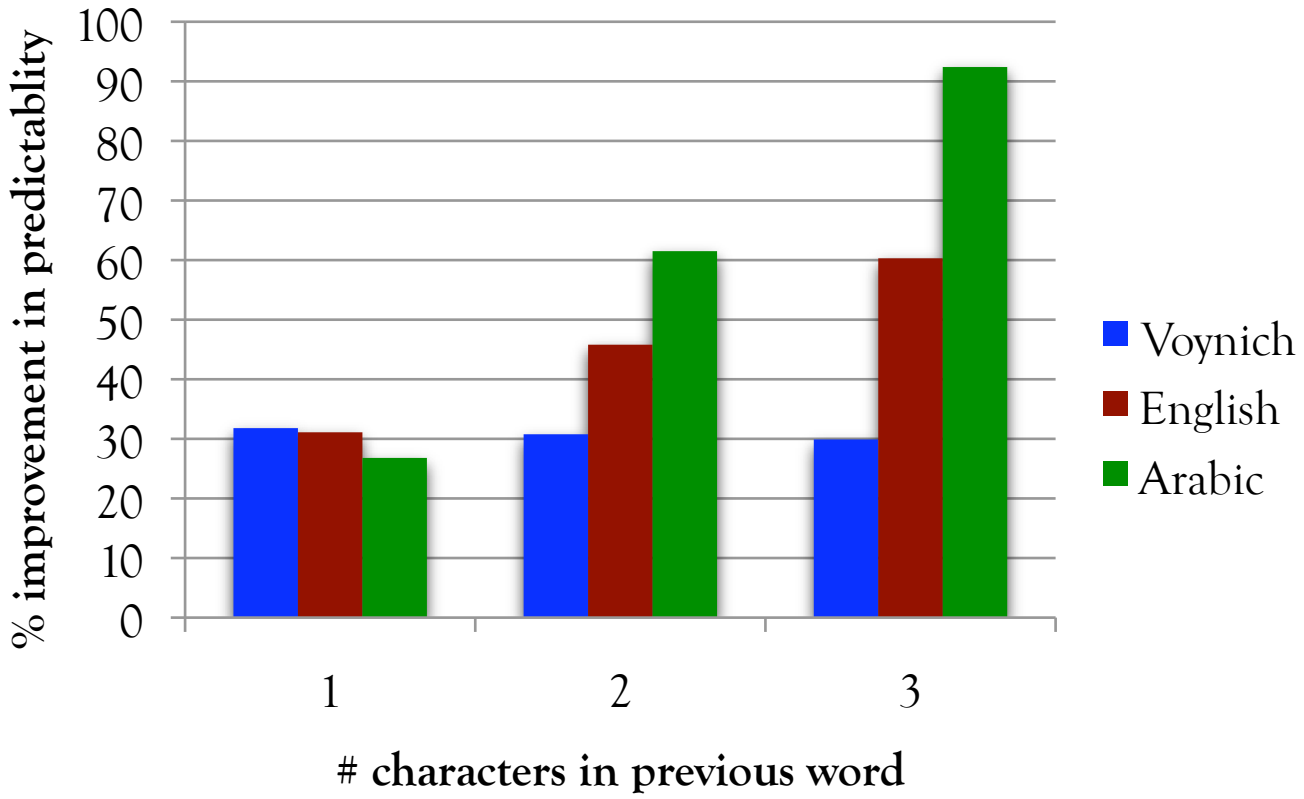
Is there Word Order?

- Very few repeated bigrams and trigrams
- Predictability of word given previous word (10-fold x-validation)



Are There Latent Word Classes?

- Predictability of first character of word improves when using last characters of previous word

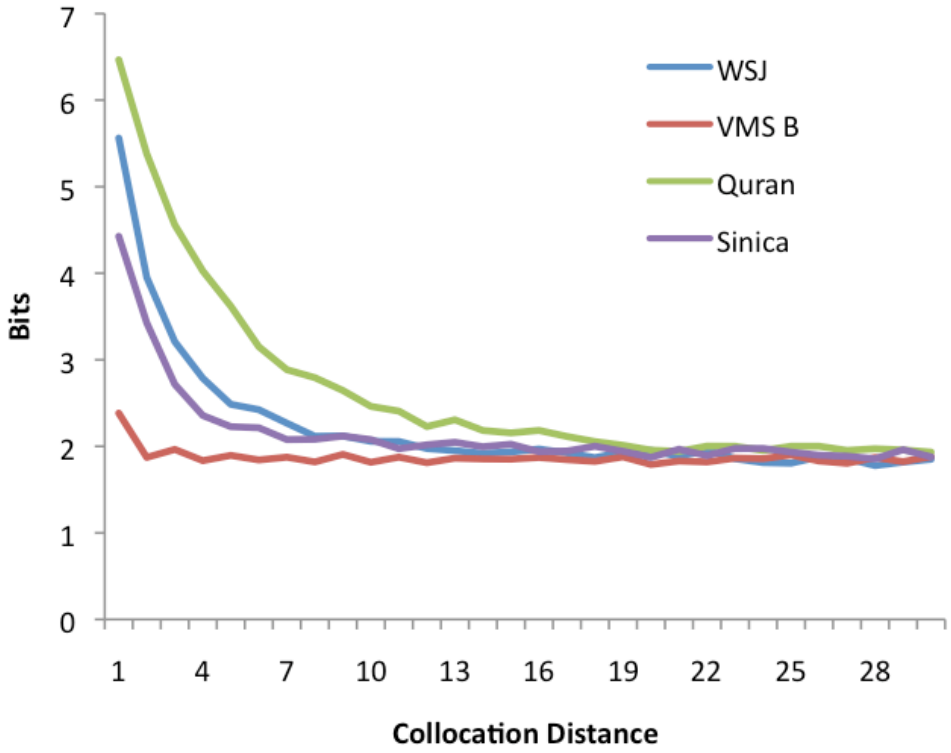


Are There Latent Word Classes?

- Clustering words using bigram HMM

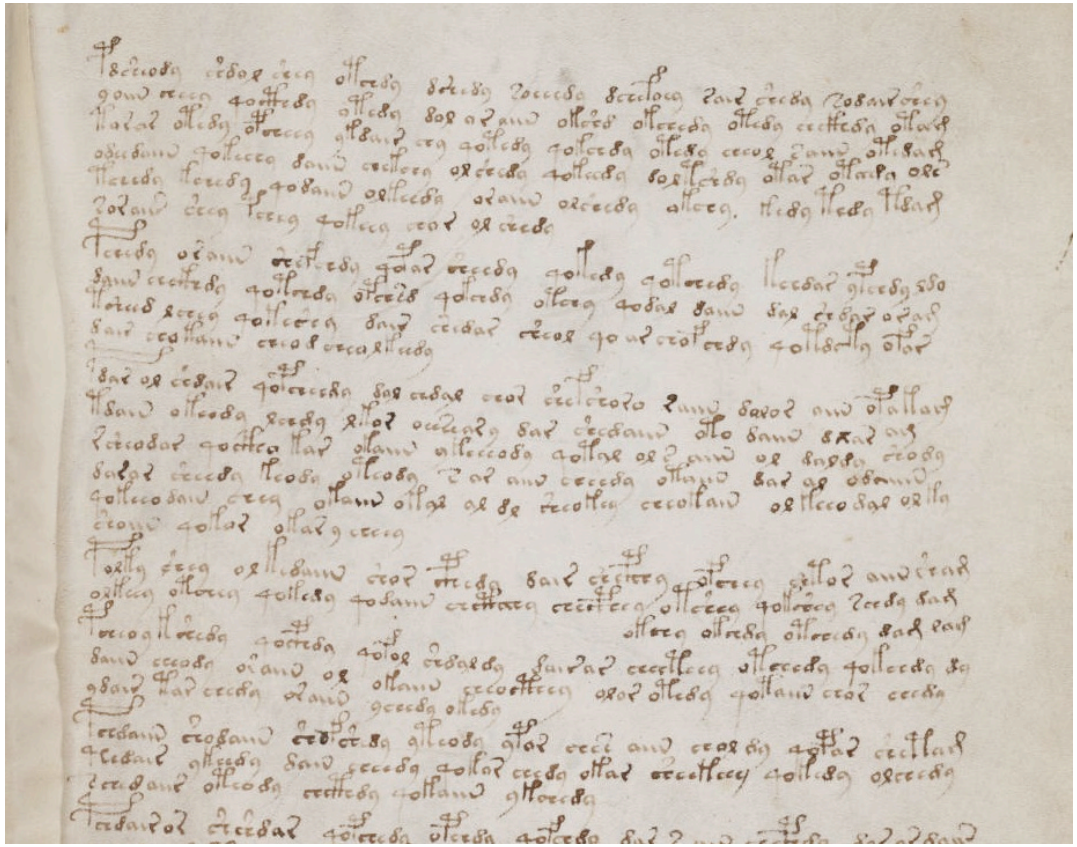
Are there Long-Distance Collocations?

- Pointwise mutual information (PMI) – degree of 'collocationness' – of word1 and word2 at distance d
- Overall collocationness C = average PMI at that distance over all word pairs

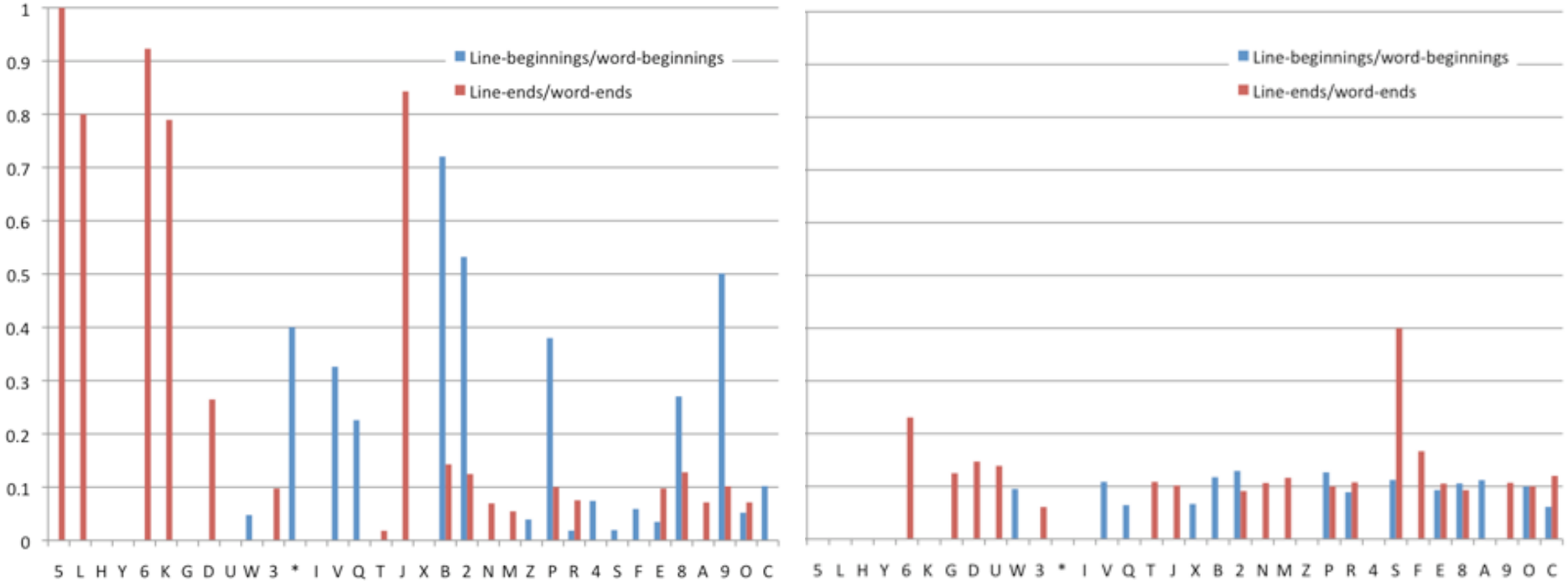


Is it Prose?

- Text is visually left to right
- But some characters occur disproportionately at line-edges



Is it Prose?



Scrambled words within lines

Do Pages have Topics?

- Spiky TF-IDF distribution

Visualization of TF-IPF values of words in a Voynich B page

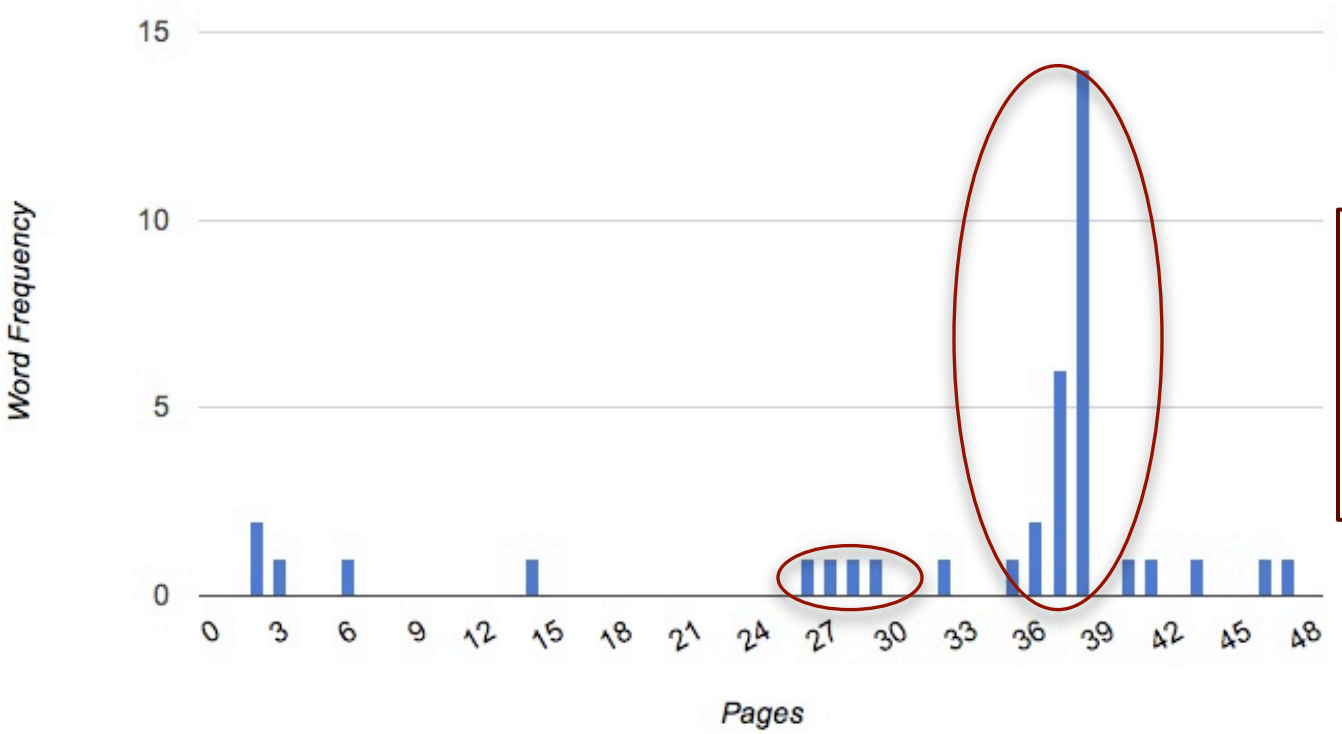
Do Pages have Topics?

- Would spikes appear if words were independent of page?

Visualization of TF-IPF values of words in same Voynich B page, where words are scrambled over the document and repaginated

Do Pages have Topics?

- Spikes across consecutive pages

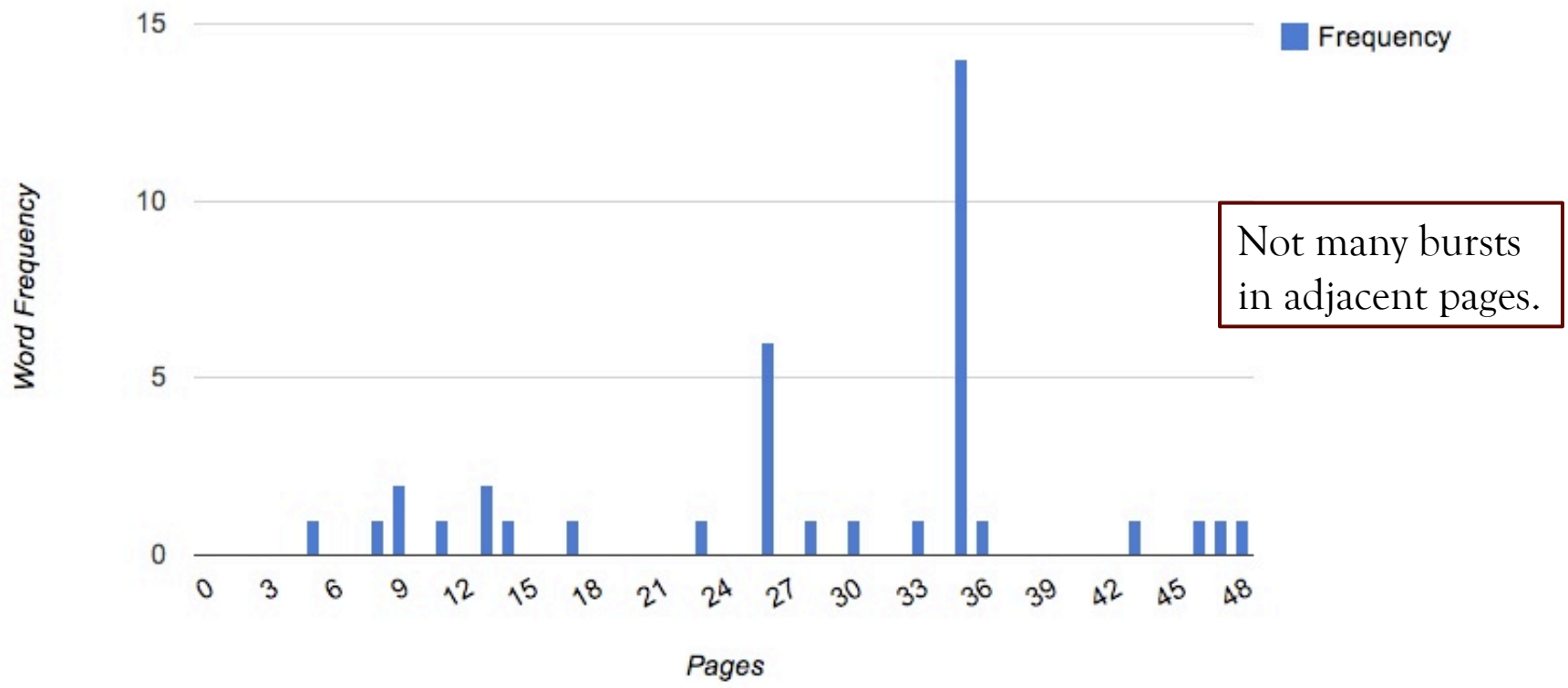


Indicates narrative, with topics bursting through a few consecutive pages

Distribution of ꝺſſccꝺ9

Are the Pages in Order?

- Look at same plot, with pages scrambled



Are the Pages in Order?

- Quantitative measure:
  % of pages P_i where most similar page P_j is adjacent to P_i

  - If pages are not independent stories + pages are in order, this number will be high

Voynich B	Voynich B – pages scrambled	English WSJ	English Genesis	Quran
38.78%	0%	1.34%	25.0%	27.5%

Strong page order

Articles are independent

Single narrative, continuity across pages

Does the Text Correspond to the Illustrations?

- Yes, within sections (recall page-similarity plot)
- Many of the 'bursty' words in a page are used next to or inside images. Are they captions/proper names?

- Without more fine-grained annotation of images, we don't know if it's true at the level of the page or line

(But how to annotate?)

Is it an Encoding of Latin/Ukrainian/Chinese?

- Several claims of decoding on the Internet
  - All use arbitrary scramblings, rearrangements, and mappings to force-fit to some message
  - This can be done for any string
- Hoax theories – propose ways of generated Voynich text from tables or FSTs
  - Do not explain difference from natural language