

WHAT WE KNOW ABOUT THE VOYNICH MANUSCRIPT

Sravana Reddy

The University of Chicago

Kevin Knight

Information Sciences Institute, USC

June 24, 2011



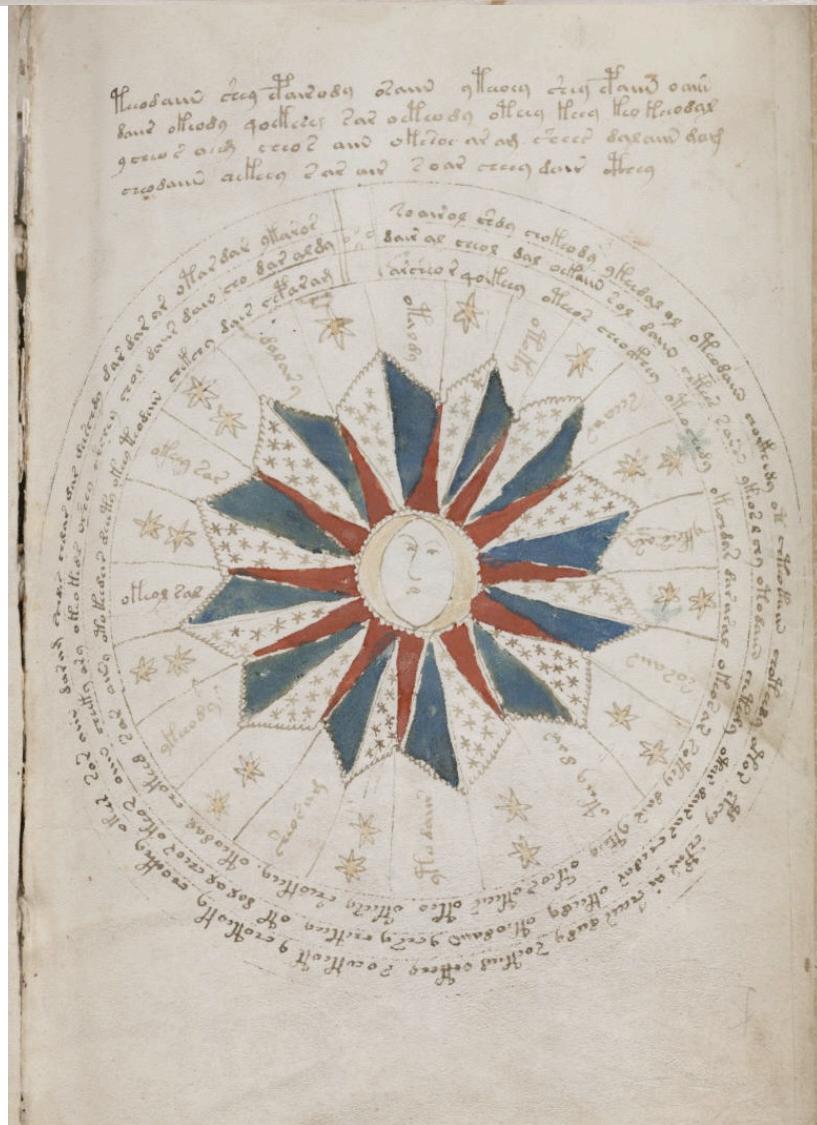
BACKGROUND

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



BACKGROUND

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



BACKGROUND

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



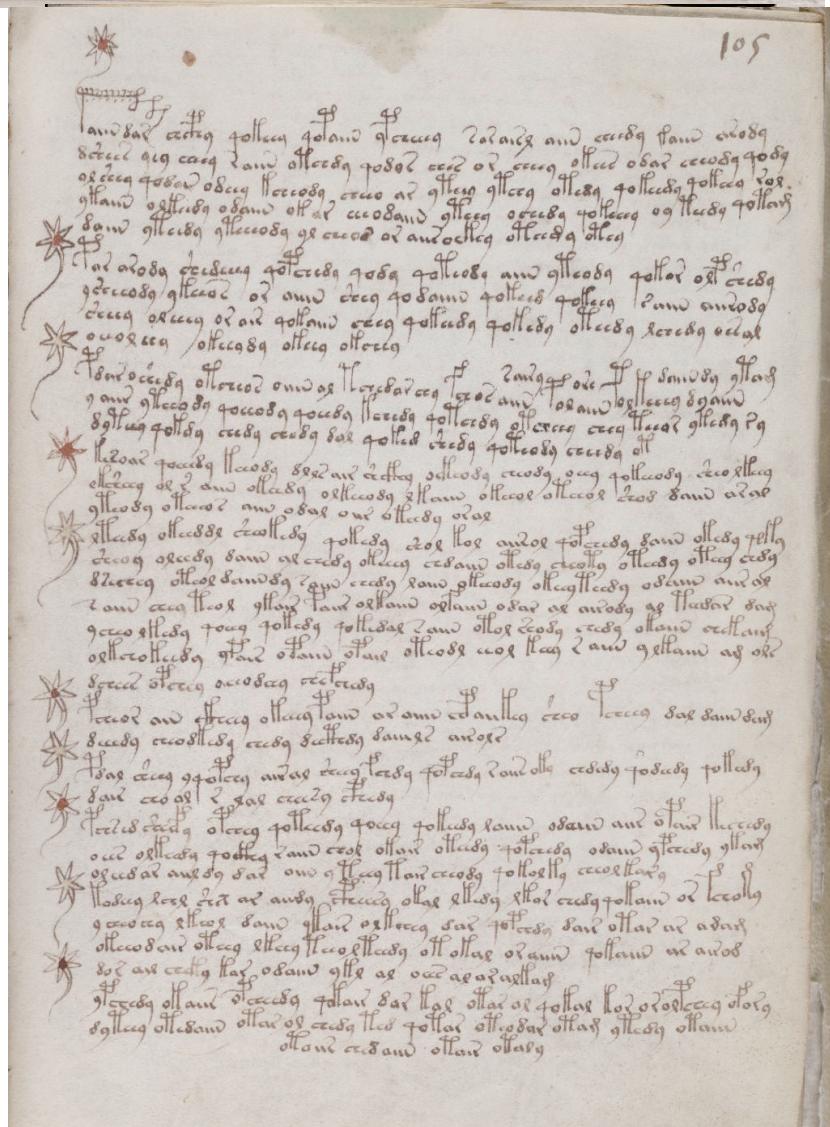
BACKGROUND

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



BACKGROUND

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



HISTORY



(Roger Bacon)



John Dee (England),
16th century



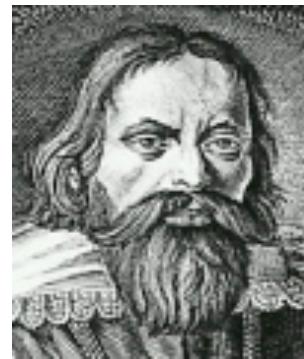
Emperor Rudolph II
(Austria), 16th century



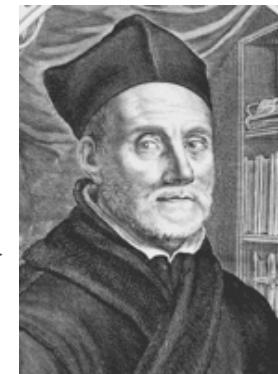
HISTORY



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

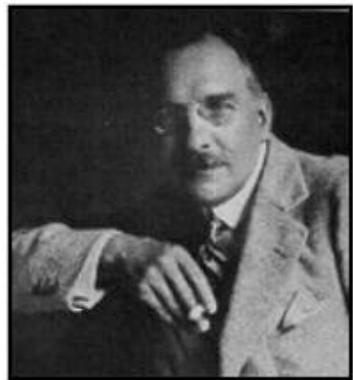


Jan Marek Marci
(Prague),
17th century



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?

Is there a
narrative?
Topics?

Is it
prose?

Does the
text
correlate
with the
illustrations?

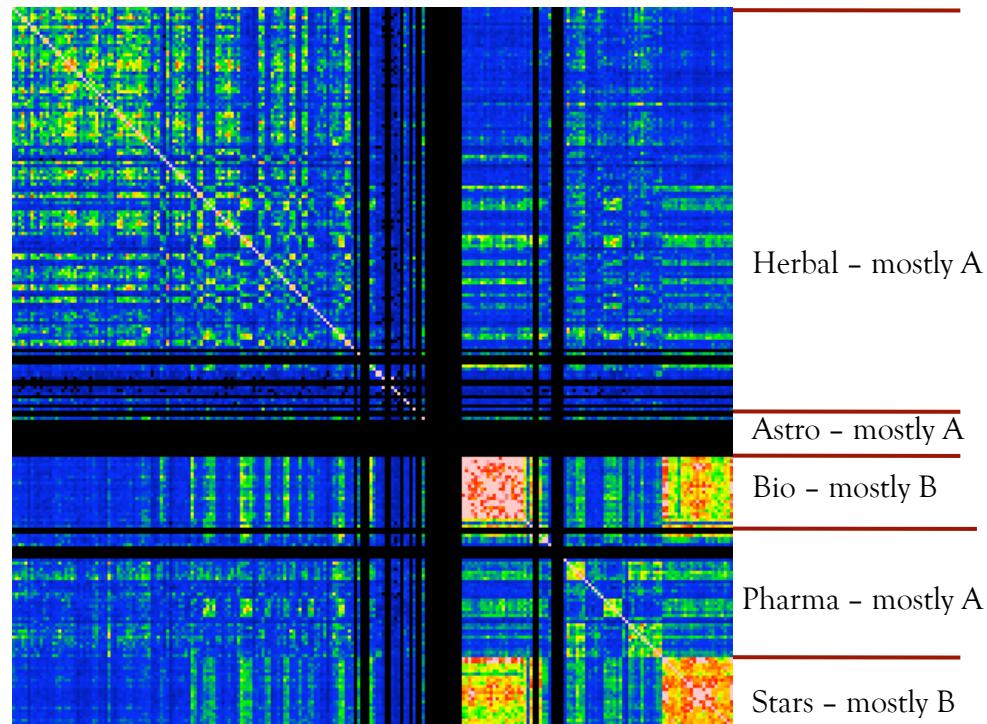
Are the
pages in
order?

How predictable
are letters within
words?

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

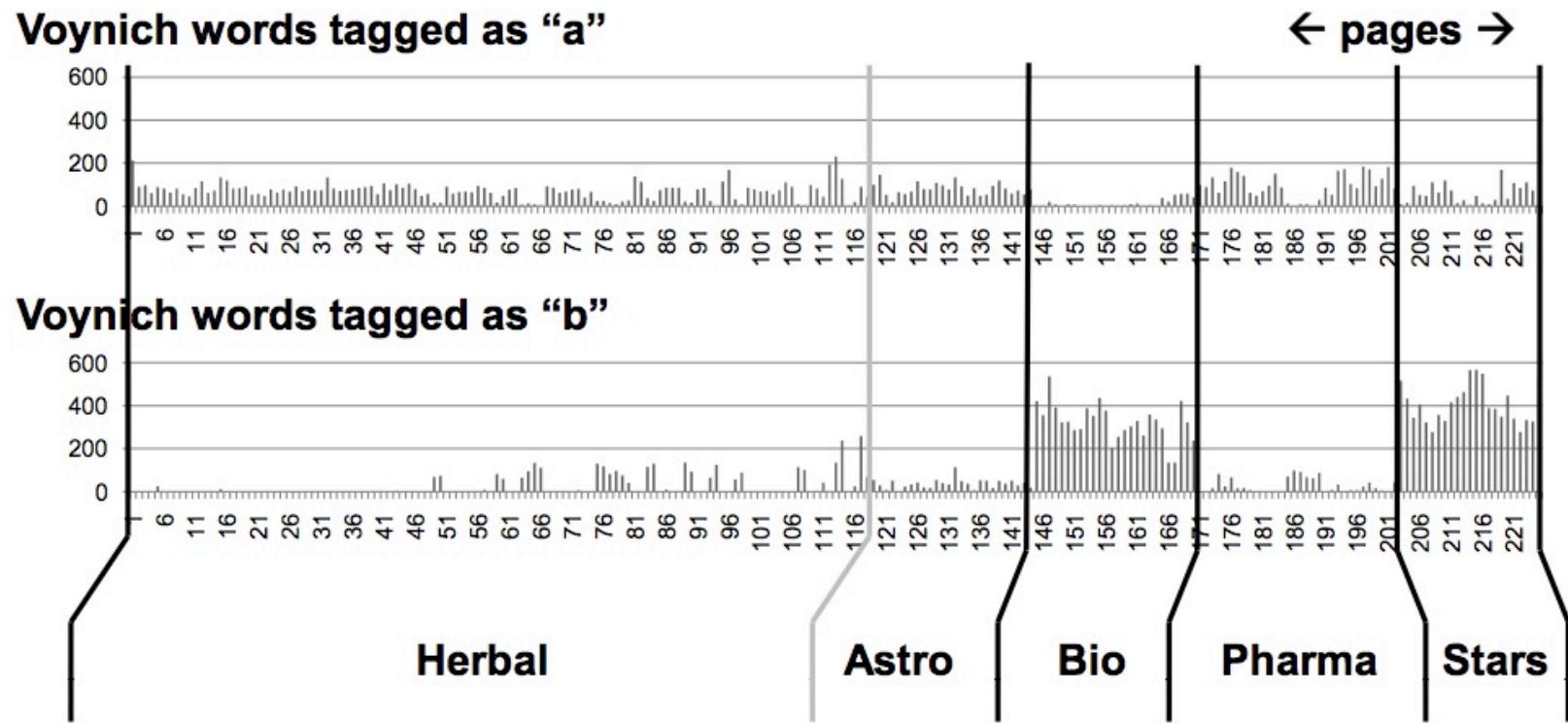
HOW MANY AUTHORS?

- 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
 $\alpha=A, \beta=B, \gamma=C, \delta=D, \epsilon=E, \mu=F, \dots \tau=4, \vartheta=8, \rho=9$
- 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

But even long words
seem to have only
one ‘vowel’!

- Characters are syllables or morphemes

Only 35 characters...

- Abjad like Semitic scripts – most vowels not written

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

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

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

Arabic with no diacritics

Most likely!

DO LETTERS HAVE CASES?

- Some characters only occur at beginnings of paragraphs/lines

ꝝꝑꝑꝑꝑꝑ

- Decorative uppercase?



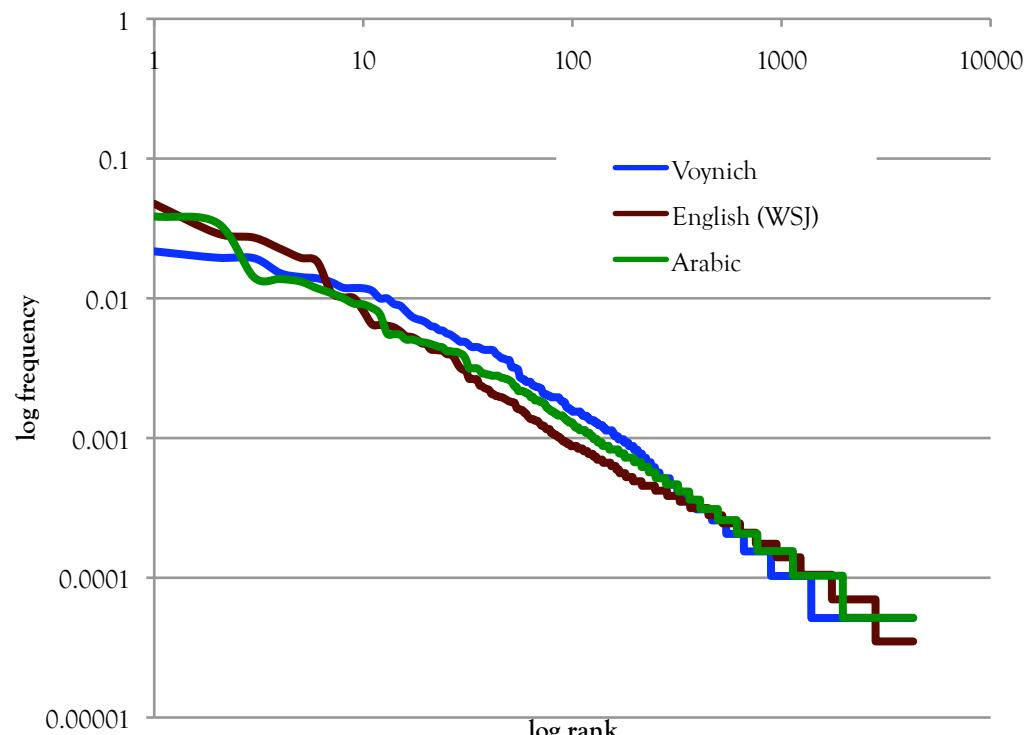
- To find lowercase equivalent of A in English:
 - replace every instance of A with another character
 - compute decrease in word entropy
 - character with highest decrease is lowercase equivalent
- But lowercase of ꝝ is ꝑ, ꝑ is ꝓ, etc.



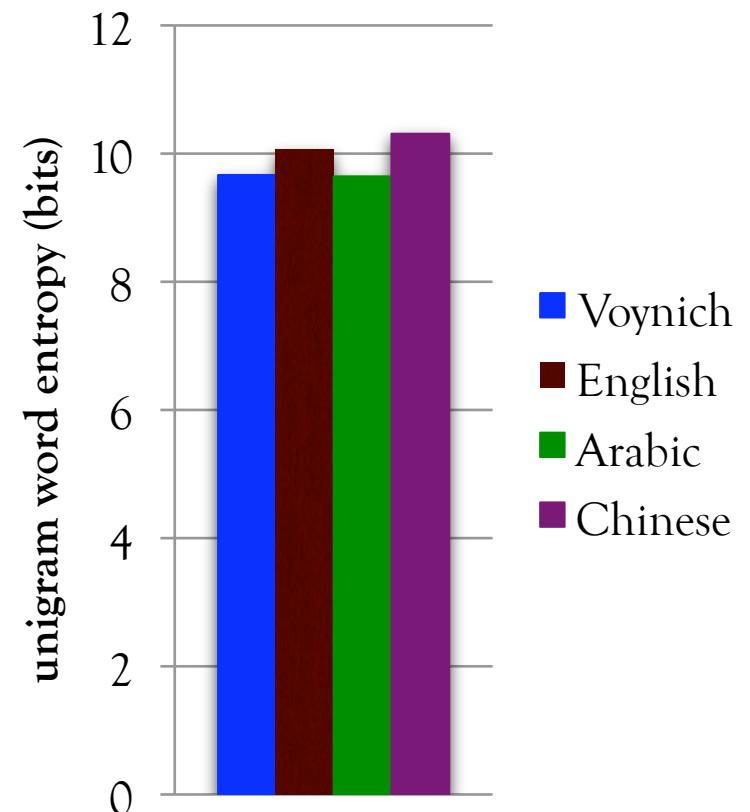
IS THERE PUNCTUATION?

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

WORD FREQUENCIES?

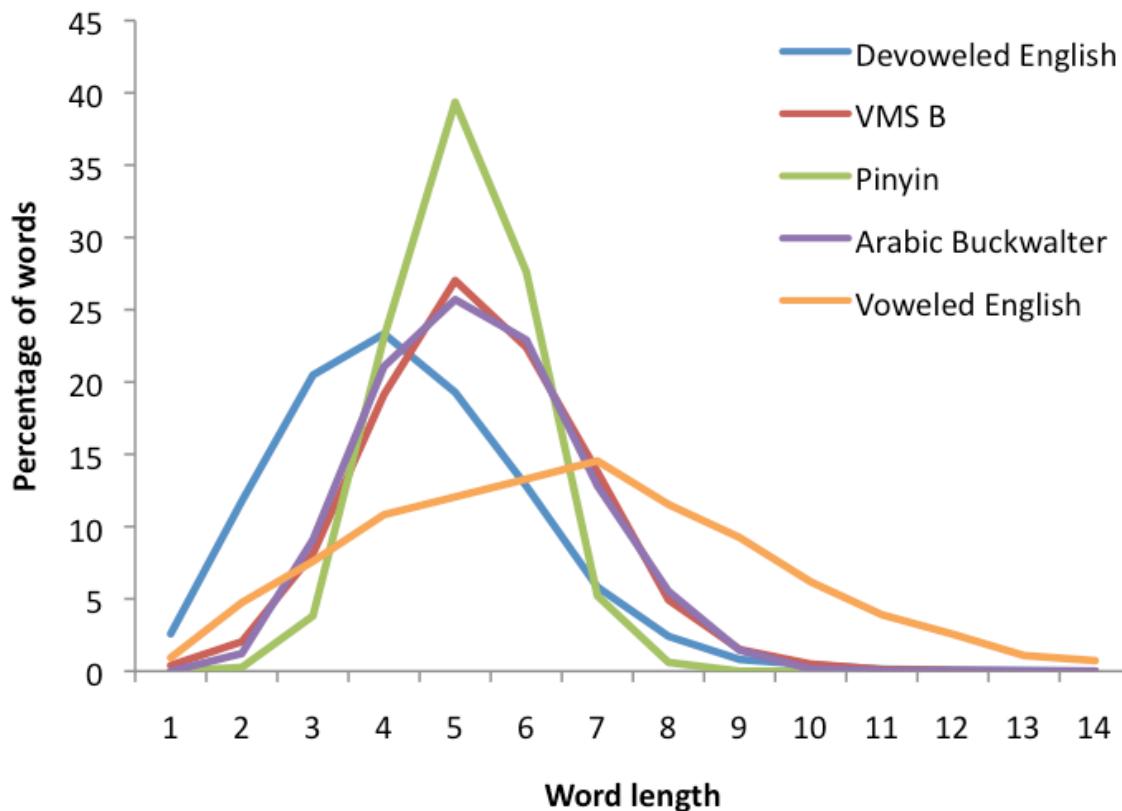


Frequency/rank distribution



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?

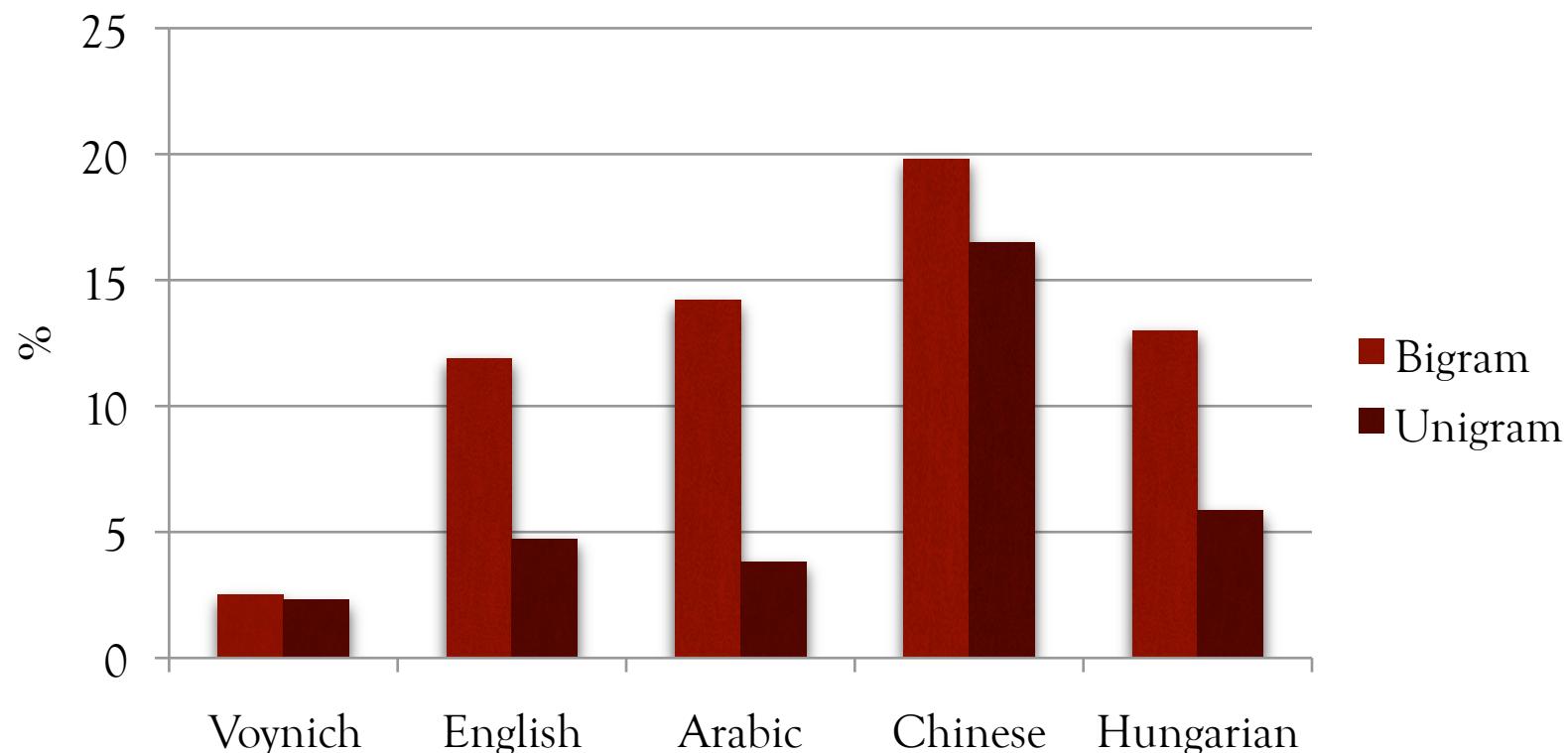
- Unsupervised morphological analyzer Linguistica (Goldsmith 2001): MDL to find prefix+stem+suffix segmentaion
 - ‘Signatures’: groups of affixes that take the same set of stems

Affixes	Stems
OP+	CAE CAM CAN CC2AE CC6 CC8 CC8AE CC8AN CC8AR CC8SC9 CC9FZ9 CCAE CCAJ CCC8SC9 CCC9 CCCO2 CCO2 CCO8AG CCO8AJ CCO8E CCOEFC9 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 ROEYC RS SBS SC9F SCBS SCFC SCOQ SOFS SOQ SOX SP W X ZCBS ZCOFC ZCQ ZFC ZFS ZX...

Stems have similar spellings

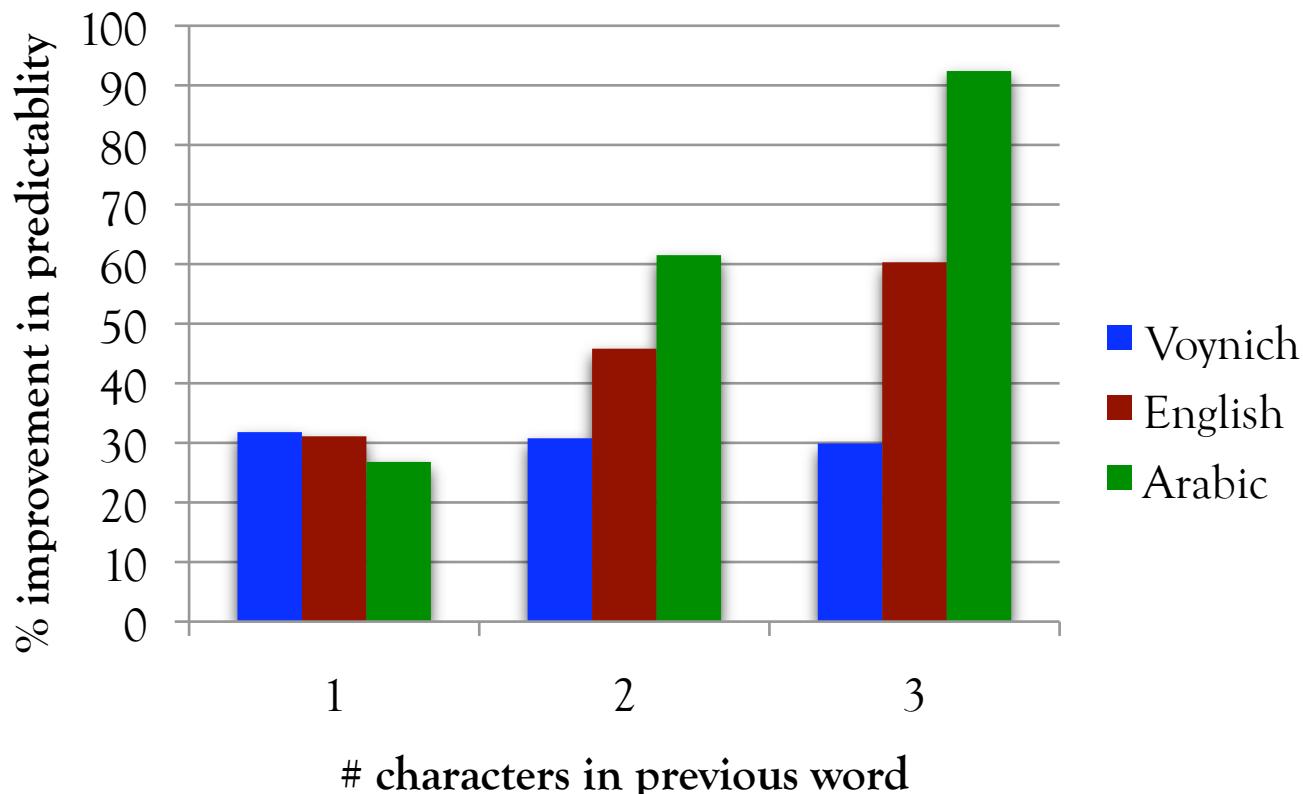
IS THERE WORD ORDER?

- 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

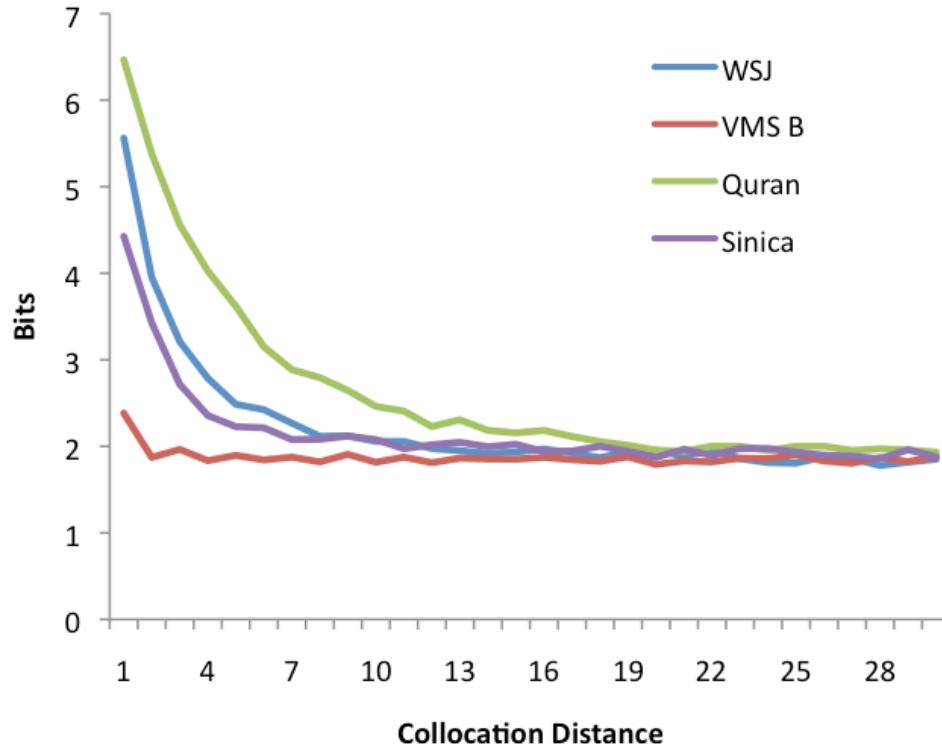
2AM 2AN 4OFAM 4OFAN 4OFC89 4OFCC89 4OP9
4OPAR 8AE **8AM** 8AN 8AR 8AT AE9 EFCC89
ESC89 OE OEF9 OEFAN OESC89 OFAE OFAM
OFAN OFCC89 OPAE OPAJ OPAM OPAR S89 **SC9** SCAE
SCC89 **SCC9** SCF9 SCOE SCOR SCQ9 SCX9 SOE
SX9 SXCB9 ZC89 ZCP9 ZOE ZX9

2AE 2OE 4OBSC89 **4OE** 4OF9 **4OFAE** 4OFAN 4OFAR
4OFC89 4OFC9 4OFCC89 **4OFCC89**
4OFCC9 4OFOE 4OP9 4OPAN **4OPC89** 4OPCC89
4OPOE 89 8AE **8AM** BAN 8AR 8OE 9FCC89 EFC89 ESC89 ESC9 EZC89
FCC89 **OE** OEFCC89 OEFCC9 OEOR OFAN OFC89 OFCC89 OFCC9 OP9
OPC89 **OPCC89** OPC9 OPOE RAM S89 S8AM **SC89** SCC89 SCC9
SOE **ZC89** ZCC89 ZCOE

2OR 4OE 4OFAM 4OPAR 4OPCC9 8ARE
EFAR EFCC9 FAE FC89 O OBSC8AM **OE** **OFAM**
OFAN OFCC9 OPAM OPCC89 OPCC9 OR
PSC89 SAR **SC89** **ZC9** ZCC89 ZCF9

ARE THERE LONG-DISTANCE COLLOCATIONS?

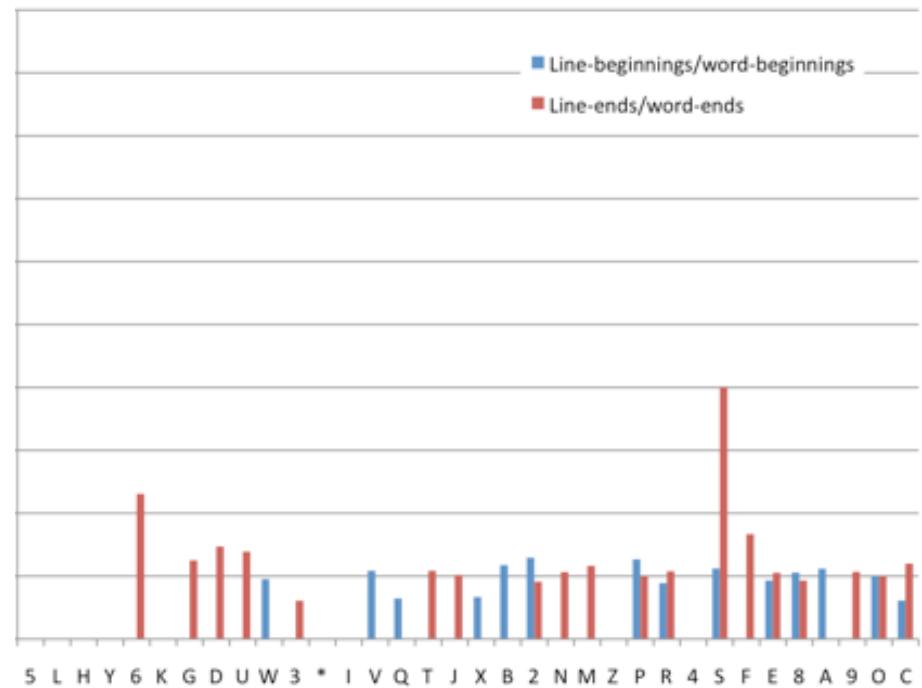
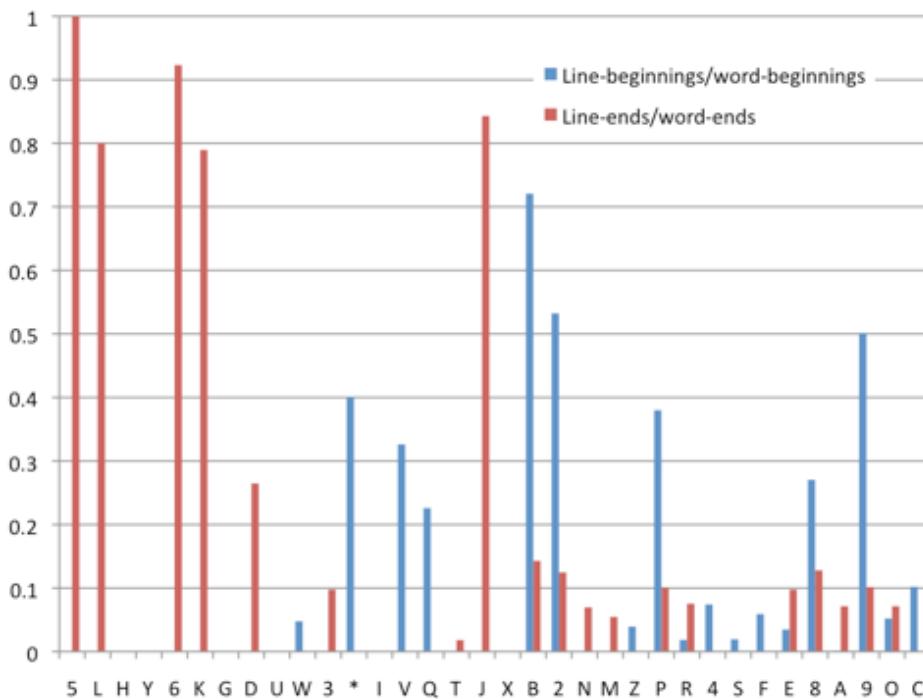
- Pointwise mutual information (PMI) – degree of ‘collocationness’ – of word1 and word2 at distance d
- Overall collocationness C = average PMI at that distance over all word pairs



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

OFCC9 ZC89 8AN ZC9 SCFCC89 ROR 4OFAN SAE FAN ZCCF9 EPAN OR 9FCC9 EZC9 FAE 4OFCC89 FAR SC9 R AN ZCC9 4OFCC9 4OF9 FCCCOR 8AR OFC9 SQ9 BSC89 SX9 EZC89 OFCC89 4OFC89 OESC89 2AE S89 OPC9 ESC89 4OFAE OFAR FCC89 2AN OPAR 2AM OPZC89 BSC8AR 8AE ZAE OPAN SAR AR ESC9 EOR AJ OEFCC9 SCQ9 ZX9 OPC89 8AT ZCQ9 E SFAE 4OFC9 SCF9 SQC89 EPC89 OFAE OP9 4OFCS89 OPCC89 4OPCC9 4OPCC89 OFC89 OPAJ 2ZC89 2 O 8AM EFAR BE ES89 FC9 SCC89 8AR9 ZCOR 4OFCOR AM ZCC9 SE FCC9 4OPAN OFAN 4OPS89 OPCC9 SXC9 EFC89 4OFCCO AE SC8AN **EFCC89** ZC8 SOE OEAN ZC8AE 4OFAM 9 SCOE EFC9 4OBSC9 OFS9 **SCAE** AEOE ZCAR ORAM 4OXC9 RAM OFC8AE SCQC9 4OPCC2 ZC8AJ OFCCC89 4OXC89 SCXC9 OPCOE 2AEFCC89 PCC89 SXC89 **PCC9** ZCFCC9 ZCCOE ZCAE ON 4OFCAR SR 8AJ EAJ 8ZCC89 9SCC9 ZAR SQ*9 EOFCC89 ER SXAE 4OPSC89 ESCOE SC8AR ORAN 4OPSC9 AT OIF*9 4OSC9 OEAFAE 8ZC9 PAR EFO **EFCC9** ZC9 AEOJ FCCOE 4OFCSC89 8SC8 4OVSC89 SO89 4OFCOE ESC8 SFCC9 8SCC9 EOFAJ 4OF OFCCOE SCBSC89 8OM ZCCX9 PC9 OPC8AR OPC2 4OFC89 OPC89 SC889 4OFCO89 OPCC8 AK EVS9 SAM PAT CCC2 OJ EFAJ **FCCC9** SCO2 8AK ZCCF EOP9 ZCFAN 4OFC89 OFCAE EFCCO89 ZC08 ZC8AN BSAE OPCAE OPCCAJ 4OAN 4O8 SCO SCOFC89 AEAJ **OFCCO** EFCSC9 EFCSC89 SCCFAN OPCCC9 EFAT ESAE OPCCOE SO FCSC89 OEFCC89 4OFCO8 4OSC89 OPCO ZCCO8AR SCOJ OPARAE OAM OEFCCO EFCCOE EFAE EFCCC9 PC8AJ OFC8AN ZOF **4OFCOE** 4OFCO2 OPCC8AN EFCCC89 SCAJ SCXC8 OSC9 FSOES8AR AIIB SCPAEZ9 4CCA8 4OFC89 ZCCFZ9 SOFSC9 SX*9 9SCC8AN 9FCC8AM RCCC9 OEA3 AIF*9 ZFAM 8ZCCO OPSC8C9 OPCOEAT 4OZCO 8SC8AR 9FCCC2 BSC8C9 OPCO80 BS8AJ OFC8AN ZCP ZCOPAJ ZPAR OESCO89 ZCQC9 ESCOCFAJ 4OFCO8OR SCQ89 4OPC8E EOC89 SC8C9 EFAK O*OR F98CC89 ZCCP SCOP989 2ZCO PSC8 FCOQC89 4OFCZC9 FCCZO89 **OFCSC89** ESR 20 AEAE O*AR 2OAM OPCO8AM 4OPCC8AM PCC8AN ZCOFAR RF9 SPAR 4OTAN ZCOPSC89 ZFC9 **4OFCAN** ZFCO89 8ZCCOPCC9 PCAR SOEFCC89 ESCS89 4OCCCO SC8A FCCO8AE PCO OFCOJ 4OFCO8 EFC8EFC9 **PCC8** 9SC8E BOEAE B9FCOR SCCV9 OBSC8AE EVSC89 ESCCOE OPCON SCAJAR 4OFCOFC89 OPCCOEFC9 EAN 4OFC89 4OFCCE SC89PCOFAN EFC8AR EFCC8AN OFCAJ PCOEFC8AN EPCCAE U OFCCC2C9 OESAE 4CCAR BOCOFCC9 PC8AN SCBSAJ 4OFCOFAN FCCE BOEFC89 SCOFCAN I*AR *AN 9ZC OFZ89 ZFCC9 SXAM

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

DO PAGES HAVE TOPICS?

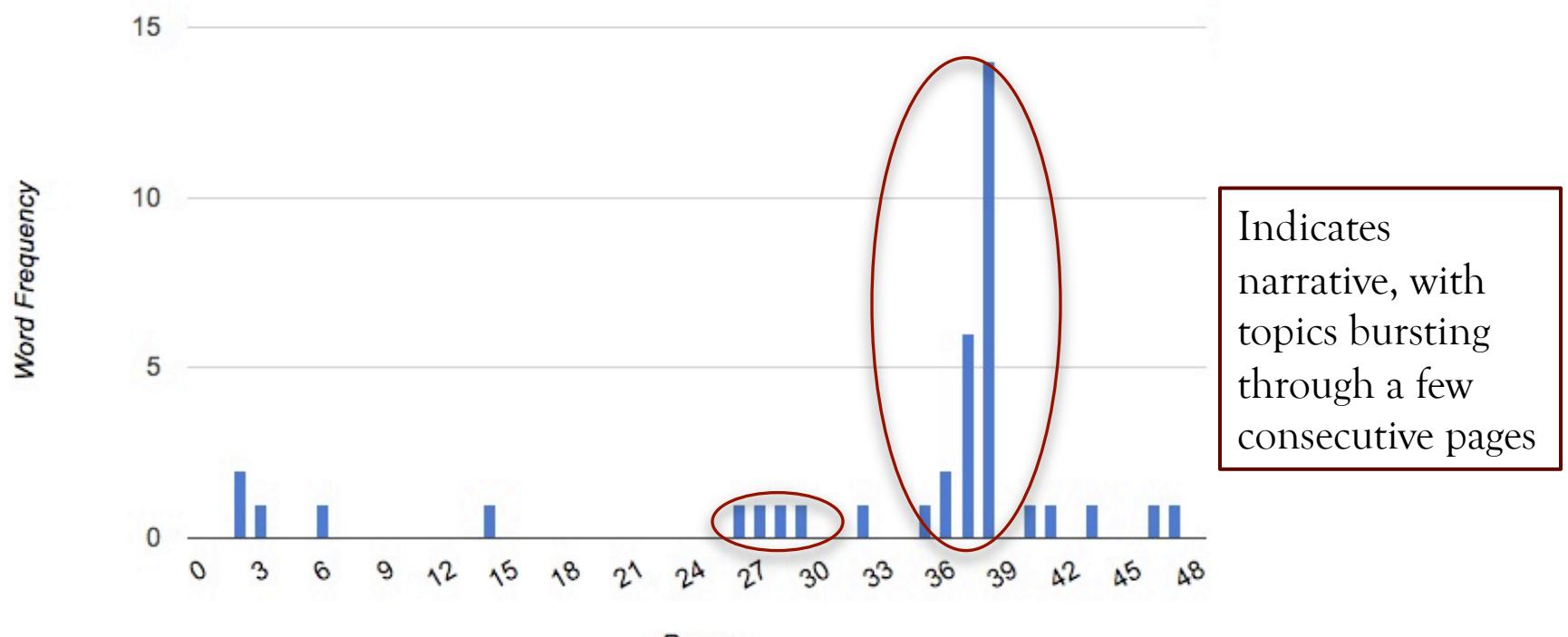
- Would spikes appear if words were independent of page?

OFAN SCC9 BAN 4OE AT OPC89 4OPAE AN ZC9 8ZC89 4OPAN AM ZQC9 PC89 QC89 AJ SCAR OFAR 2OE ZCC9 OFAM 4OPCC89 8AR 8AE OPC9 EFAM 4OFS9 AR9 OR 4OFOR SX9 OFCC89 4OF9 4OFAM OPAM ZCC89 OEOF AE 4OFC9 FAM 4OFC89 OPOE ZCO8 EO 4OPCO89 A3 OP9 OF9 4OPC89 SC9 OFCCC89 89 R9 OFS9 SO89 OFCC9 RAN OB9 OPAJ 4O 2AM 9FCC89 BSC9 4OPSC89 ZX9 8AM ZCF9 EFCC89 8AT 4OP9 OBAM SOE ZCOE OPCC9 OFAE 4OFOE EAR 4OPAM S8AR 9FAR 4OE9 SCO 2AR S89 SE SCOE OPCC89 S9 SCO89 4OFAR EAM OFC9 SCC89 OBAM ZOE FCC8AE OPAR SC2 FSC89 4CC89 2AN FC89 OPS89 OEZC9 SCB9 OEFAN 8SC89 Z9 ESC89 8SCOE OPAI2 ZCX9 OBOR SCCF9 ZO8AM ZC8AM OVSC89 SCC9 4OFE 9PAR 4OFS89 FAN OFC8AR SR SAE ZC8AE OESC89 OESC9 RAM ZCO PAR 4OFC89 4OFC89 OFCOE OE9 ROE 08AR 4OFCSC89 PAM OPCOE AU 4OBSC89 FAR S08AM OFCCO OEFCC89 SCO2 ESC9 ESCC9 4OFCOE 9FCC9 SOR OPZC89 O9 PSC89 AESC89 OEFCC2 S80E ZAE 9POR OFCCO89 4OFC89 8A3 CCC2 OFSOR BAM OBS89 4OBSC9 SC8 OR9 PAE SCF9 ZC2 9ZC9 ZP9 4OPO89 4OSC9 SOFC89 EFE ORAR EFCC9 OJ AEFAN SOAM OPS9 02AM PAT 4OESC89 8ZCC9 SEAR OFS9 RAJ SC8AM QQ89 FZ8 SCFCC89 FCC89 OFZC9 OFCCC9 4OPZC89 ZO SFAN OPCO 4OPAJ 4OPCS9 9SC8AR E89 ZCP9 OEAJ 2 SQ*9 4OFC9 ORAJ 9BSC89 ESOR SPOE 9PAE FE ORSC9 8AROJ 4OVS89 OPCAE S8OR BS8AJ SXAE 4OBSCC9 98AM ORAE ZCFAN 9FCCC89 OFCCZ9 FC9 SC9F9 FS89 OESCC9 EFCC9 9SCC89 4OFC8AM EFO OPSC2 OEFSC89 EZCOE OBZC89 ZOIF*9 ZCFAE 4OFAE89 ZCW9 EPCC9 ESCOR 9SCCO 8SCC89 4OC8AM OPATOR 4OEZC9 EFCZ89 PSC8AM SCO9 ESCOCFAJ CCO8AM POEFAE 9ZCAE OEOF OEOF 4OSCOE 4OAT 4COFCO89 9FARAN 89S9 AEZ9 9FSO OFC9P9 4OFA3 4OFAK 4OPO9 WOS9 9BO8AM 4OFCCAR 8A8AJ 4OFS289 PARAT 2SCO SCOE89 OFSAM Z29 SBAE 2AE8AJ EFSCOE ZCSOE 9Z8AN OPSCCO 4OPCCO WCO BSCC9 8A12 E8AR AROPCC9 4OPT 4OFE889 AR9E9 OEZ9 89PZC89 9SC8AN 4OFCZ89 4CFAE CC2C9 ZCBSC89 OCC2AM ZOOR 4OQCO89 SAR9 FSC RF9 9SCOEF 4OFC89 4OFAO9 8AEAE 4OEFCC9 EEOR9 POEZ9 4OEAN SCQ89 4OPO2 BSCAJ ZCCO8CC2 4OCCS89 4OFC8AT ORSQ89 OFCOXC9 SCCAE 9FCO8AM 8ATOE 9PCC8AR 8CC9 2SC9 EZCO89 EO*C89 OFSCCV9 RAE8E RCCC9 89AT S8AT PORAN 2O8AE OZCC89 OBZC9 BORAE OROR9 ZOX9 8ARS9 SCPAN 8OESCC89 QQQR 4OXC8 AFIAIF9 AEO8AR ZOY9

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

DO PAGES HAVE TOPICS?

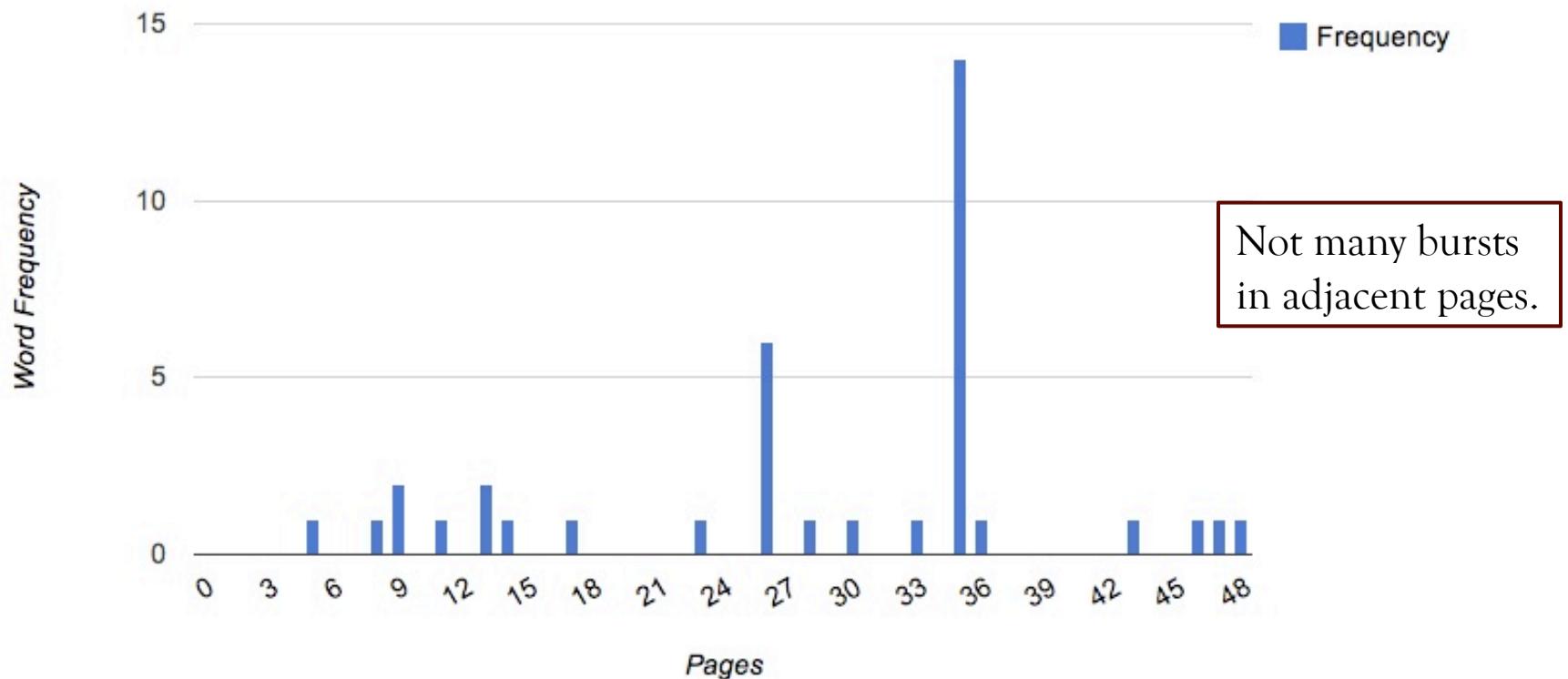
- Spikes across consecutive pages



Distribution of *allcc89*

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

မြန်မာစာ နှုတ်

