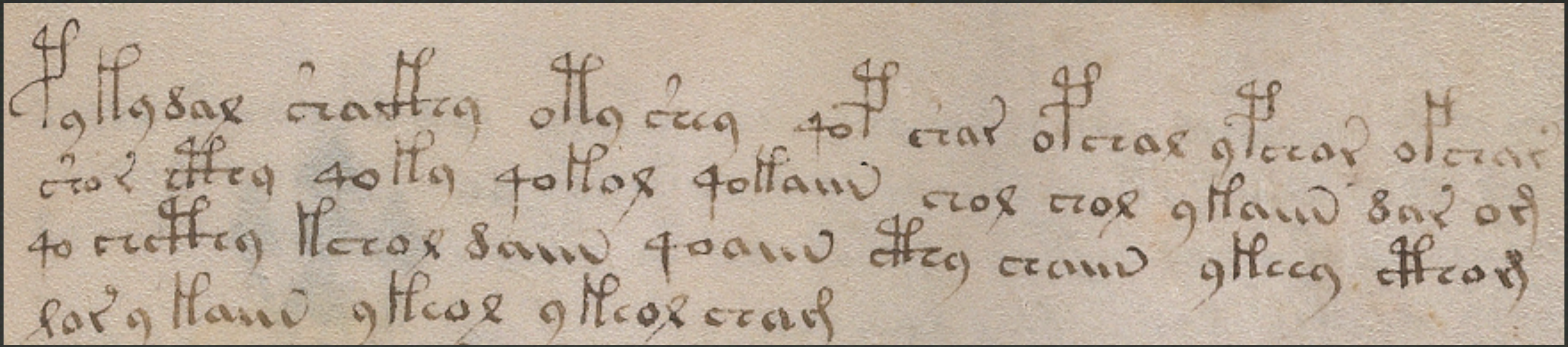
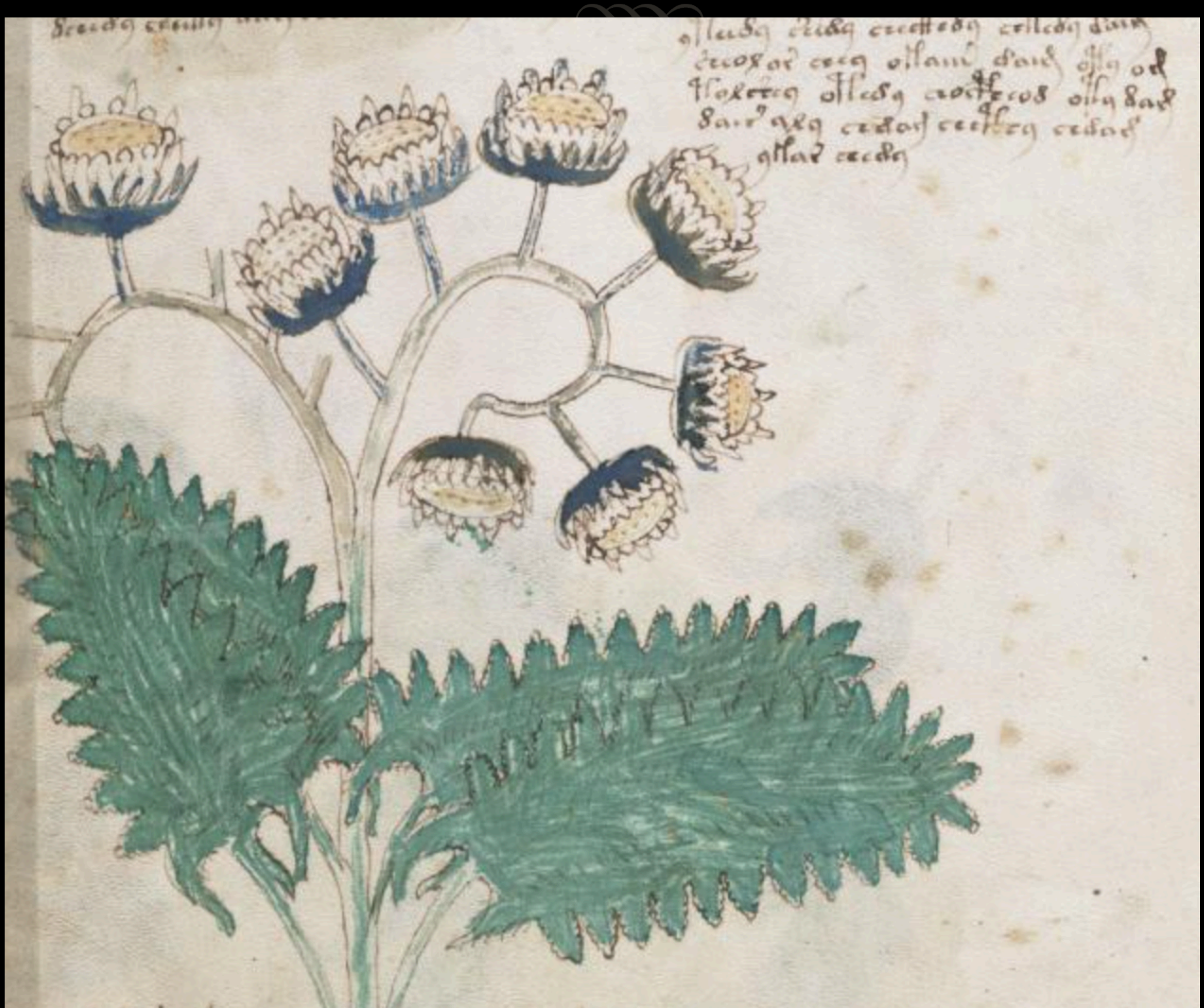
Entropy and (Mis)transcription

in an Undeciphered Medieval Manuscript

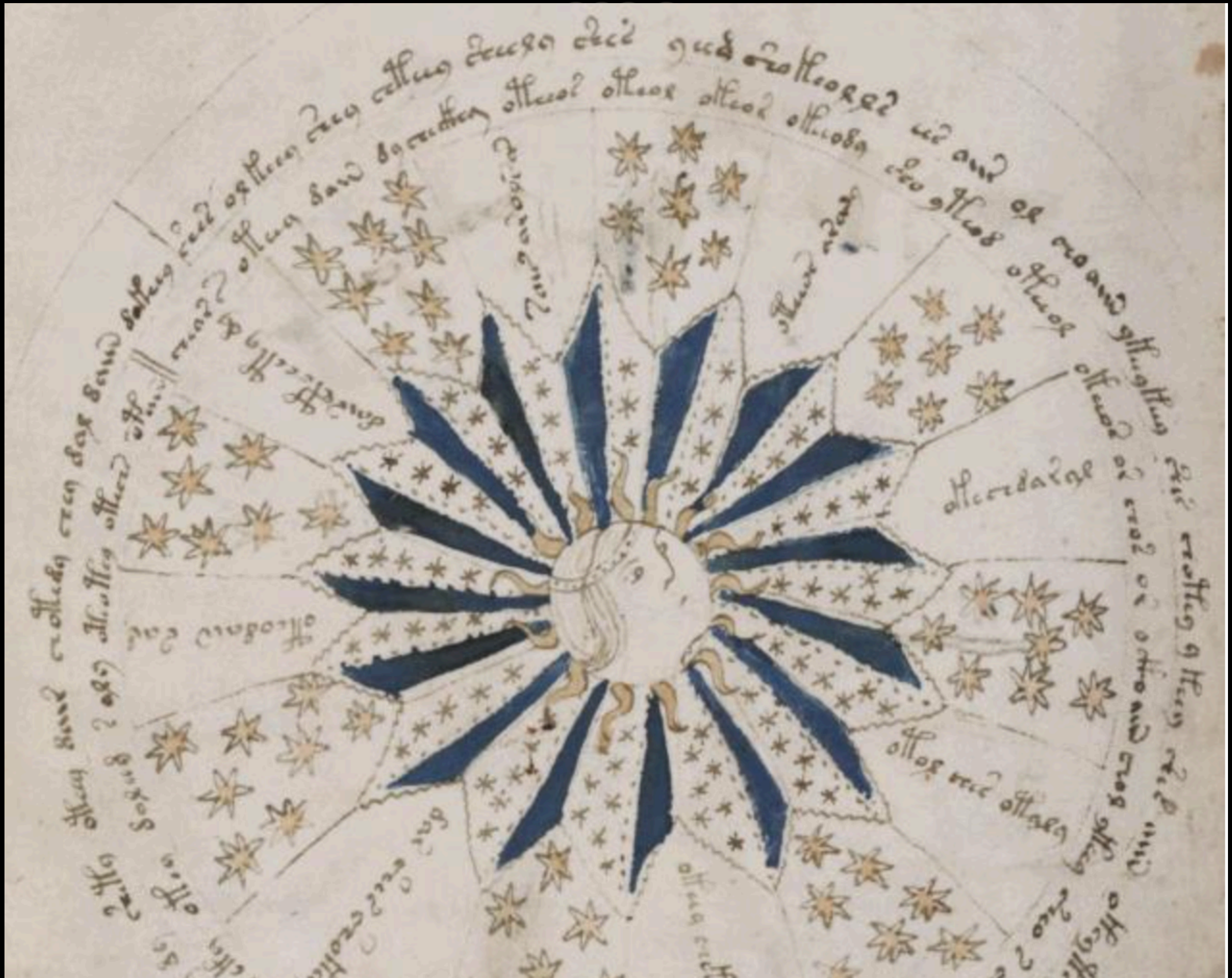
Luke Lindemann
SYNC 2018 Conference
Stony Brook University



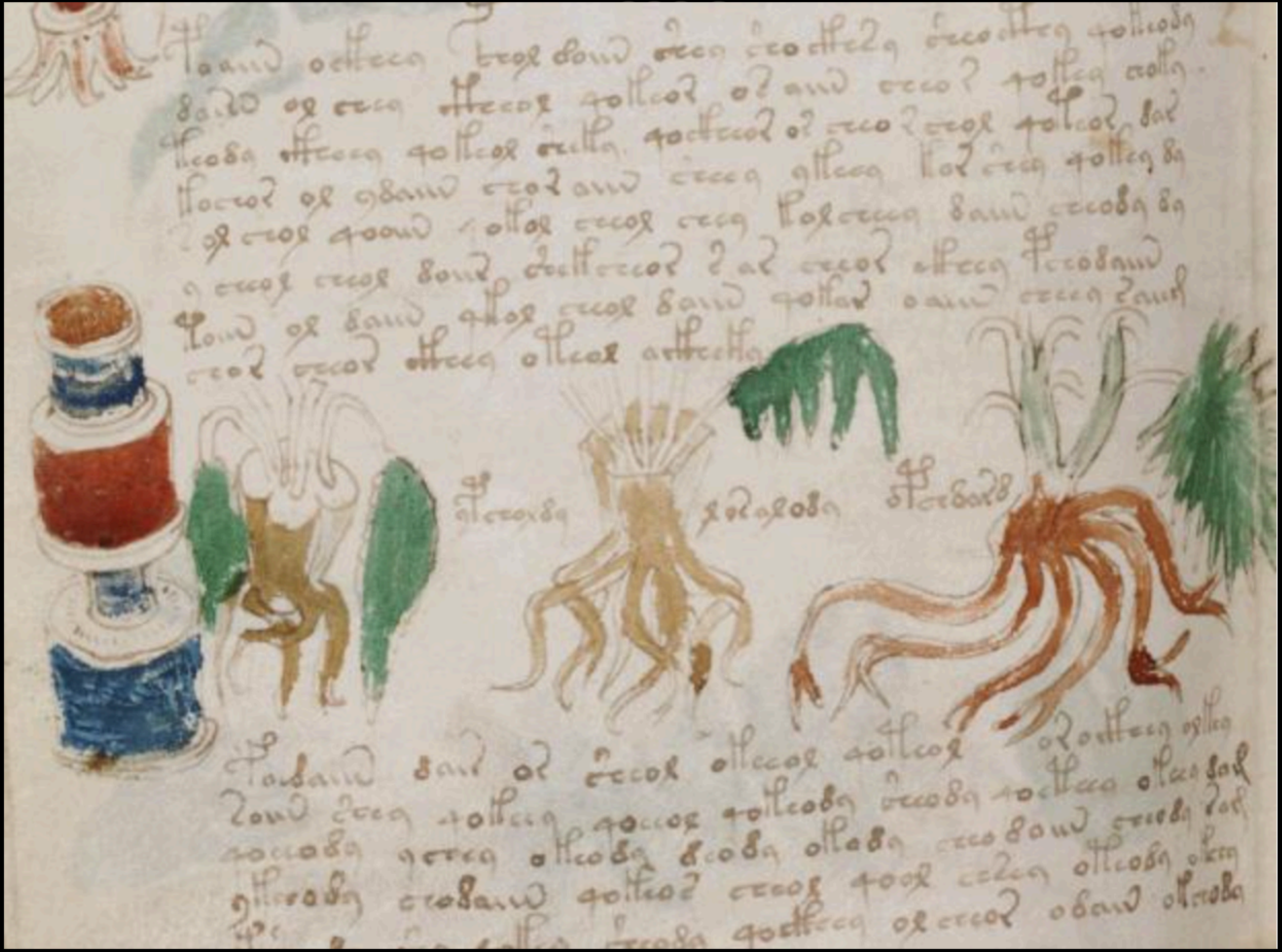
The Voynich Manuscript (Beinecke VMS 408)



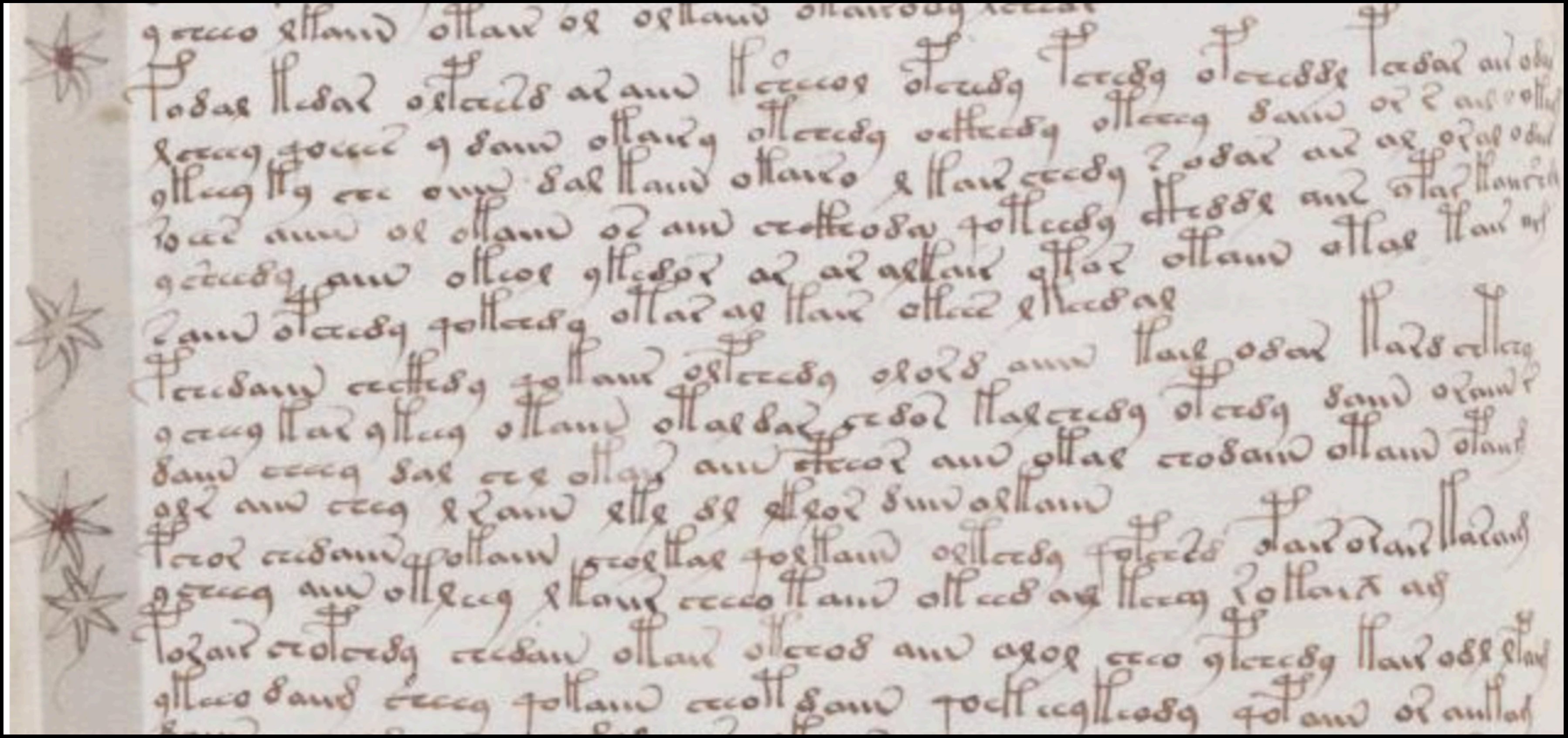
The Voynich Manuscript (Beinecke VMS 408)



The Voynich Manuscript (Beinecke VMS 408)



The Voynich Manuscript (Beinecke VMS 408)



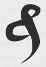
Transcription of the Voynich Text

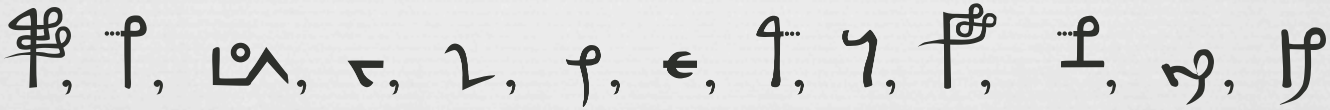
- No consensus on which marks represent a single glyph; the size of the inventory

- General agreement that there are at least two "languages" with slightly different frequency distributions: A and B.

- Six major systems: Friedman's First Study Group (**FSG**), **Bennett** (Bennett), **Currier** (Currier), **Frogguy** (Guy), Zandbergen and Landini's Extensible Voynich Alphabet (**EVA**), **V101** (Glen Claston)

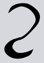
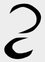
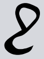
Voynich Transcription Systems

- Inclusion of rare and super-rare characters:
  -  and each occur less than 100 times in the text
  - The following glyphs occur less than 10 times each:

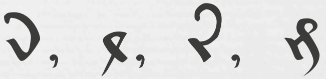


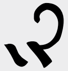
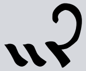
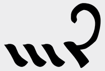
Voynich Transcription Systems

- (Minor) differences in letter variants:

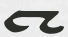
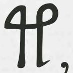
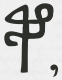
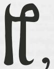
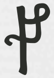
Character	EVA Transcription	V101 Transcription
	s	s
	s	t
	s	T

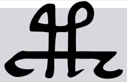
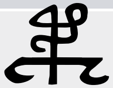
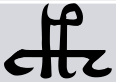
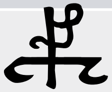
Voynich Transcription Systems

- Biggest difference: Analyzability of glyphs
  - I-sequences and end characters: 

Character	Currier Transcription	EVA Transcription
	I	i
	T	ir
	U	iir
	0	iiir

Voynich Transcription Systems

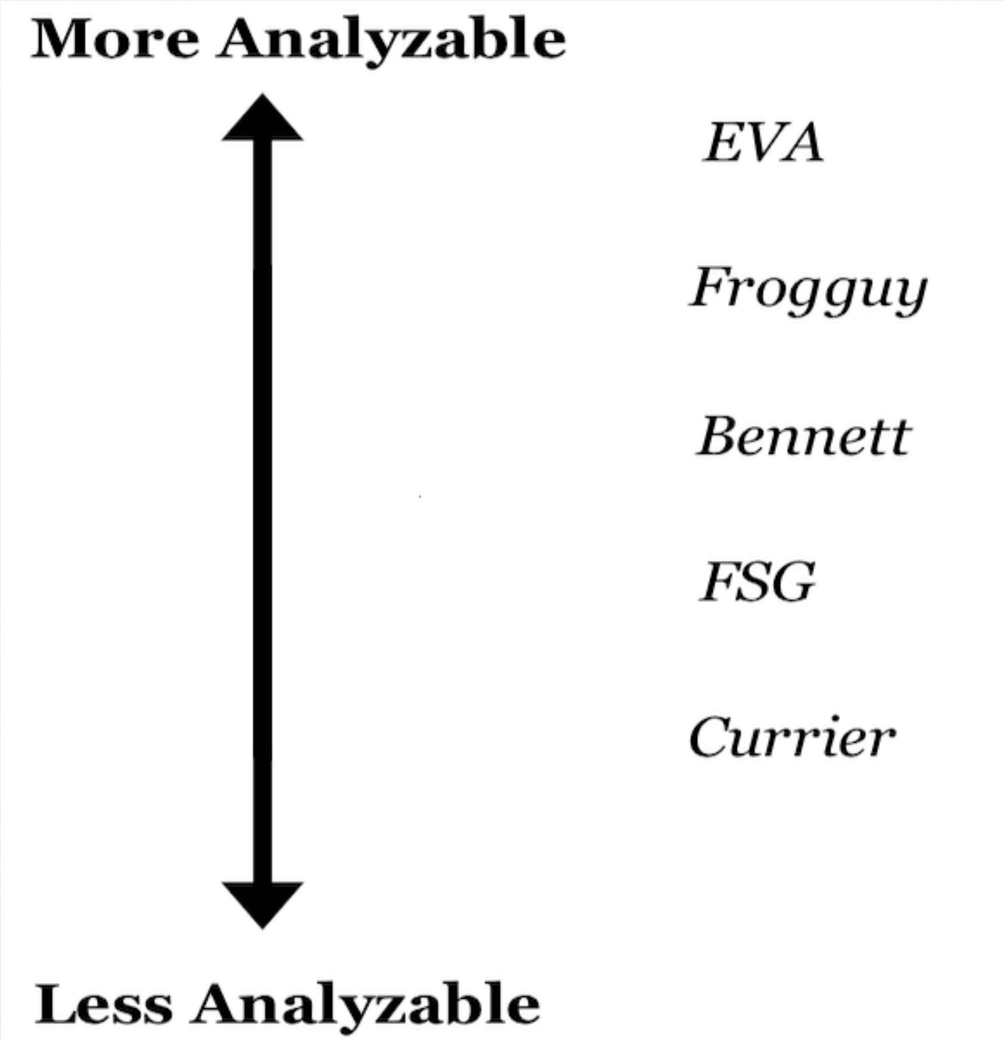
- Biggest difference: Analyzability of glyphs
  - Bench () and Gallows (, , , )

Character	Currier Transcription	EVA Transcription
	Q	cTh
	W	cPh
	X	cKh
	Y	cFh

Analyzability of Transcription Systems

- The EVA is designed to be convertible to other transcription systems like FSG and Frogguy.
- I take it to be the **minimally-analyzable** transcription: the smallest possible units are letters
- In a **maximally-analyzable** transcription multiple units make up a single letter
- Currier's transcription system is close:

Analyzability of Transcription Systems



More Analyzable

EVA

Frogguy

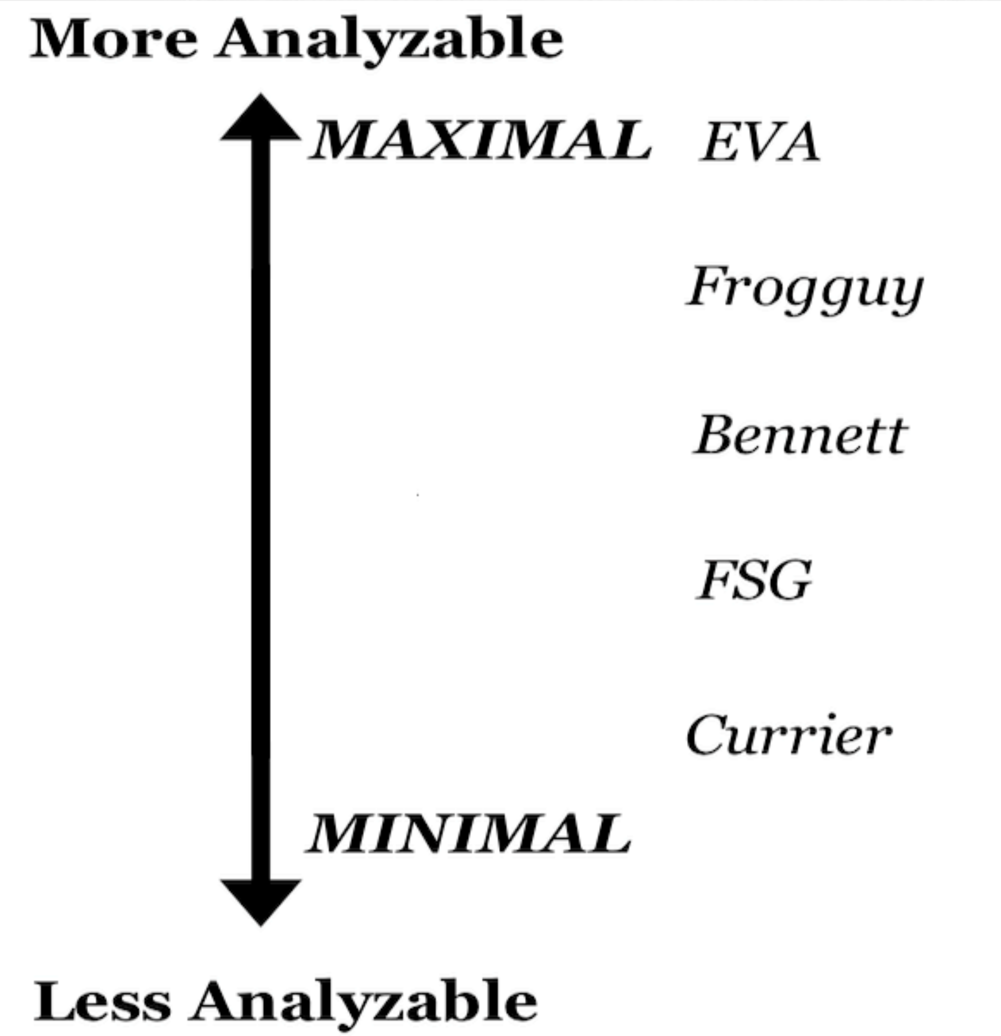
Bennett

FSG

Currier

Less Analyzable

Analyzability of Transcription Systems



More Analyzable

MAXIMAL *EVA*

Frogguy

Bennett

FSG

Currier

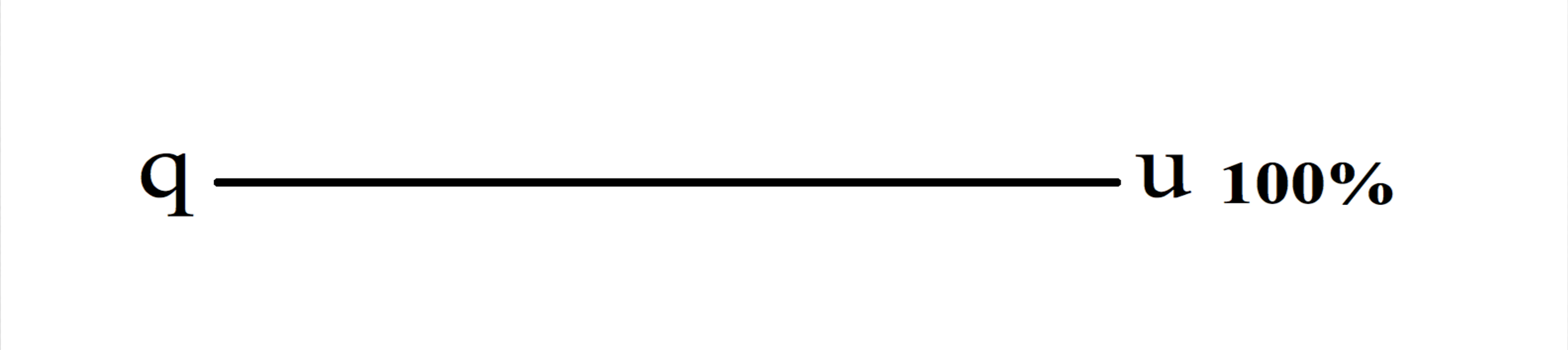
MINIMAL

Less Analyzable

Character Entropy

- **Conditional character entropy:** can be thought of as the **overall predictability** of the letters in a text.
- Given a particular letter in the text, how easy is it to predict what the next letter will be?

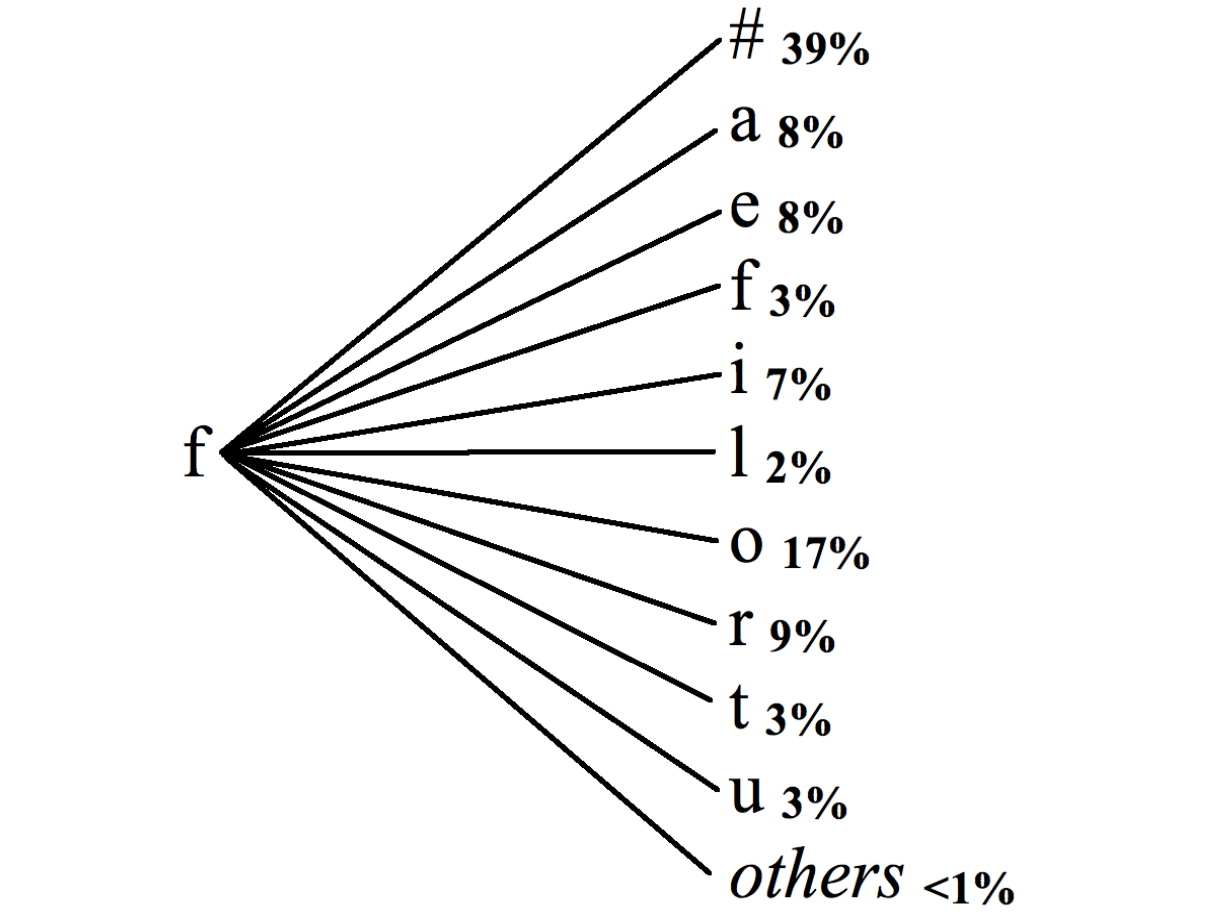
English Conditional Character Probabilities*



q u 100%

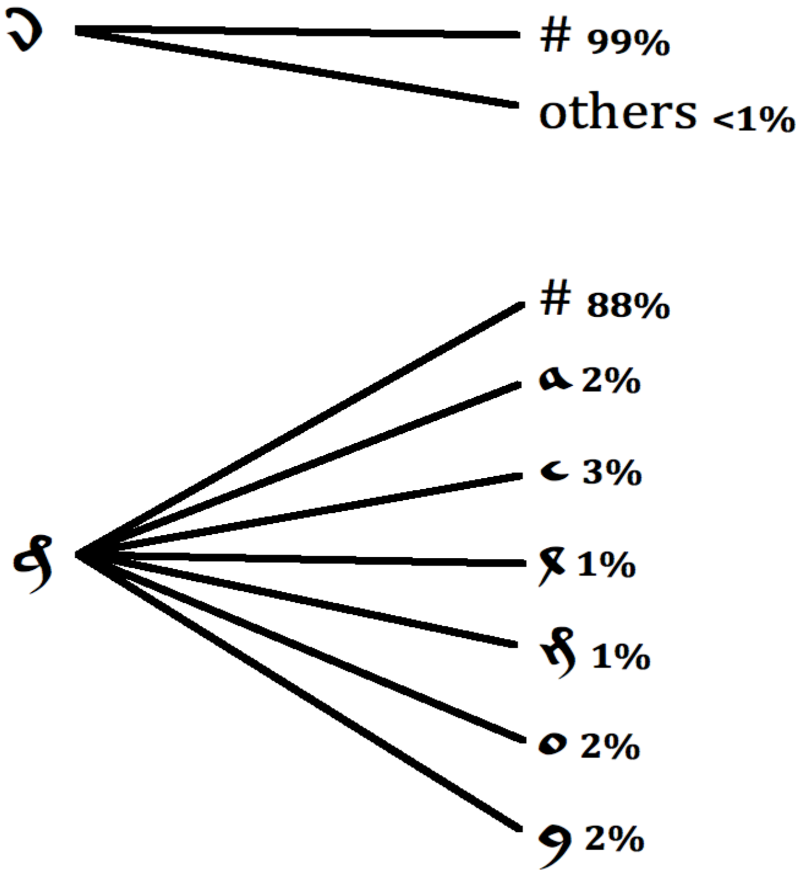
*Compiled from Doyle's *The Hound of the Baskervilles*

English Conditional Character Probabilities*



*Compiled from Doyle's *The Hound of the Baskervilles*

Voynich (EVA) Conditional Character Probabilities



99%

others <1%

88%

2%

3%

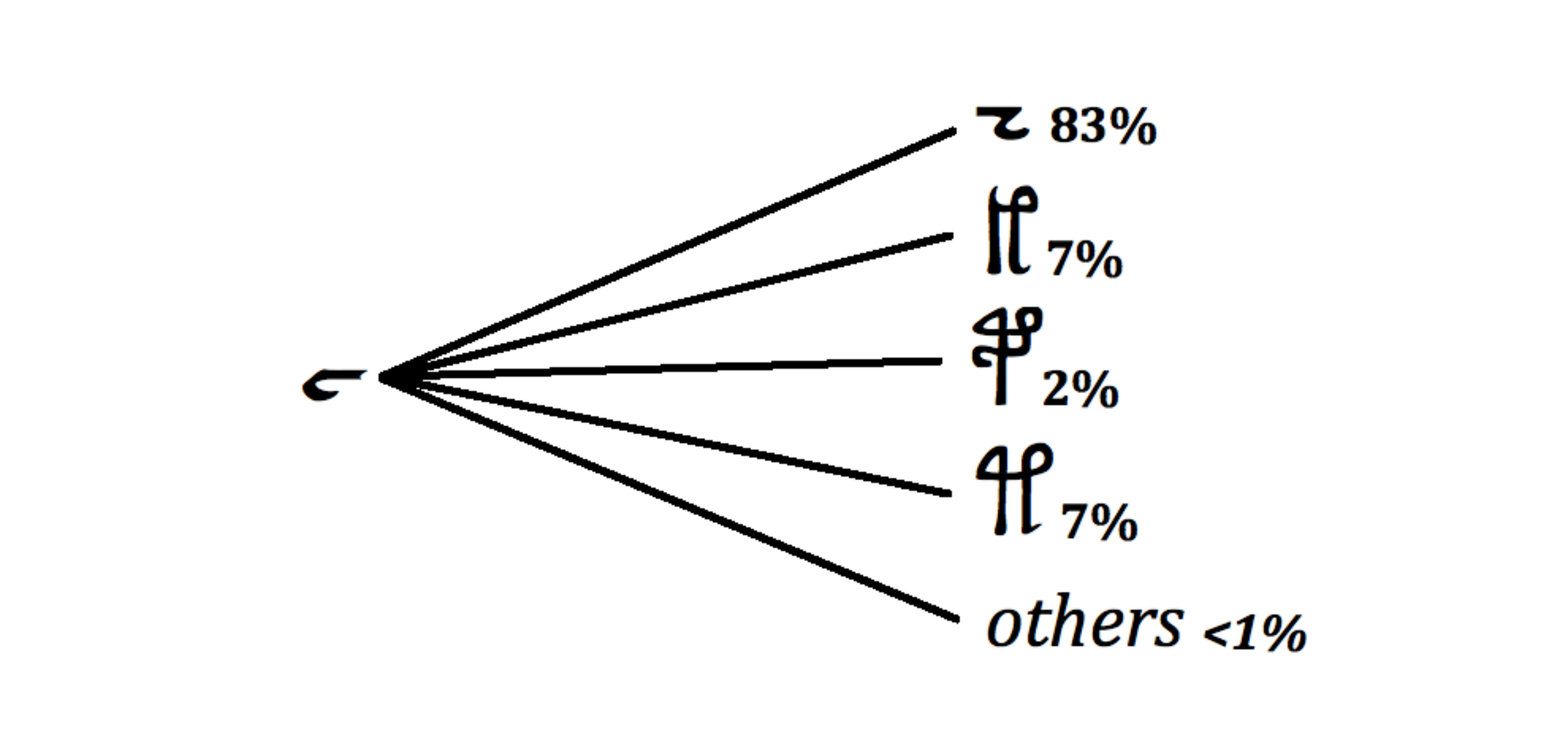
1%

1%

2%

2%

Voynich (EVA) Conditional Character Probabilities



Conditional Character Entropy

- Second-order conditional entropy *(h2)*
- Summed probabilities of a character given the previous character, weighted by the bigram probability:

$$H(X|Y) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{P(y_j)}{P(x_i, y_j)}$$

- Equivalent to absolute bigram entropy minus absolute character entropy: h2 – h1 = *h2*

Conditional Character Entropy

- Bennett (1978) notes that the conditional entropy of Voynich (*h2*) is surprisingly low
- This means that Voynich letters are unusually predictive
- Bennett compared the *h2* value of Voynich to Hawaiian

Conditional Character Entropy

Language	# Characters	h2
English (Shakespeare)	28	3.308
German (Wiese)	28	3.337
French (Baudelaire)	28	3.14
Latin (Julius Caesar)	28	3.27
Hawaiian (newspaper)	13	**2.22**
Voynich (Bennett)	22	**2.454**

Adapted from Bennett (1978)

Conditional Character Entropy

- Stallings (1998): transcription plays a big role in the information entropy values

- Number of characters in the alphabet makes a big difference (contra Bennett)

Language	# Characters	h2
Hawaiian (newspaper)	13	2.454
Hawaiian (phonemic)	19	2.650
Voynich H-A (Currier)	33	2.313
Voynich H-A (FSG)	24	2.286
Voynich H-A (EVA)	21	1.990
Voynich H-A (Frogguy)	21	1.882

Hypothesis

- Known European texts have an *h2* range ~3.0-3.4 while Voynich has an *h2* range ~1.8-2.4
- The low *h2* values of Voynich are due to properties of the script and the ways in which it has been transcribed.
- Investigation of *h2* values in different texts can tell us about script conventions as well as point to the likelihood of transcription errors.

Currier Language and Entropy

Language	Length (words)	# Characters	h2
Voynich (EVA)	41,368	22	2.200
Voynich A (EVA)	12,100	21	2.180
Voynich B (EVA)	25,688	22	2.073

Transcription and Entropy

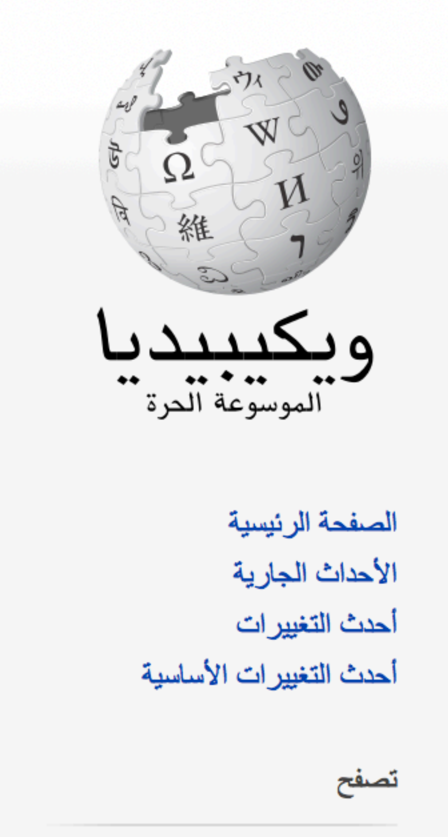
Transcription	# Characters	h2
Minimal (EVA)	21	2.200
Maximal	37	2.448

Somewhat higher *h2*, but still not in the 3-3.5 range

Abjad Hypothesis

- Reddy and Knight (2011) note that certain statistical properties of the text more closely resemble abjads, in which only consonants are written.
- This could plausibly explain the difference in *h2*, especially if there are certain character forms for the ends or beginnings of words (as in Arabic)
- The main (partial) abjads in use today are Arabic, Hebrew, and Syriac

Abjads and **Syllabaries**

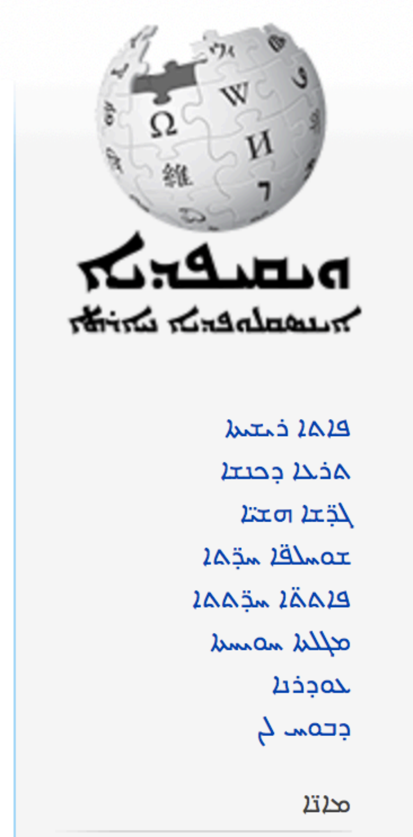


ويكيبيديا
الموسوعة الحرة

الصفحة الرئيسية
الأحداث الجارية
أحدث التغييرات
أحدث التغييرات الأساسية

تصفح

Arabic

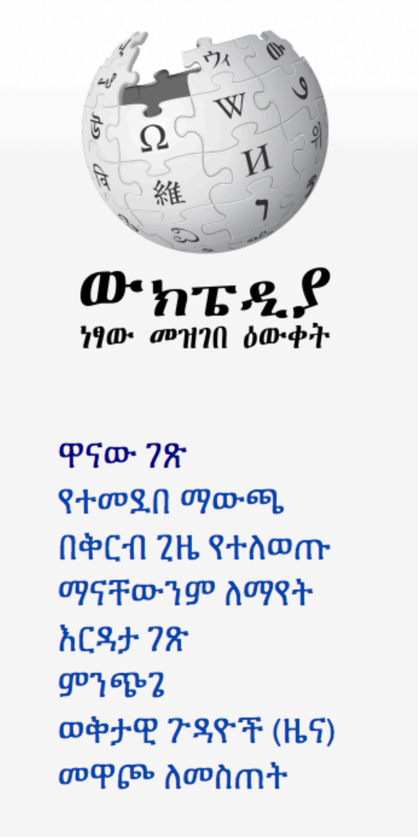


ܘܝܩܝܦܕܝܐ
ܐܝܢܣܩܠܘܦܕܝܐ ܚܐܪܬܐ

ܦܐܬܐ ܪܫܝܬܐ
ܬܪܥܐ ܕܟܢܫܐ
ܓܕ̈ܫܐ ܗܫܝܐ
ܫܘܚܠܦ̈ܐ ܚܕ̈ܬܐ
ܦܐܬ̈ܐ ܚܕ̈ܬܬܐ
ܡܠܠܬ ܚܘܫܫܬ
ܥܘܕܪܢܐ
ܕܥܘܣ ܠ

ܡܛܐܬܐ

Syriac



ውክፔዲያ
ነፃው መዝገበ ዕውቀት

ዋናው ገጽ
የተመደበ ማውጫ
በቅርብ ጊዜ የተለወጡ
ማናቸውንም ለማየት
እርዳታ ገጽ
ምንጭጌ
ወቅታዊ ጉዳዮች (ዜና)
መዋጮ ለመስጠት

Amharic

Abjad Entropy: Hebrew

Language	# Characters	Size (words)	h2
Ancient Hebrew (Bereshit)	28	19,334	3.553
Ancient Hebrew with Vowel Marking	42	19,334	3.317
Medieval Hebrew (Maimonides)	28	28,303	3.554

Slightly higher…

Abjad Entropy: Arabic and Syriac

Language	# Characters	Size (words)	h2
Arabic (500 wiki pages)	51	1,130,958	3.718
Syriac (all wiki pages)	27	25,992	3.522

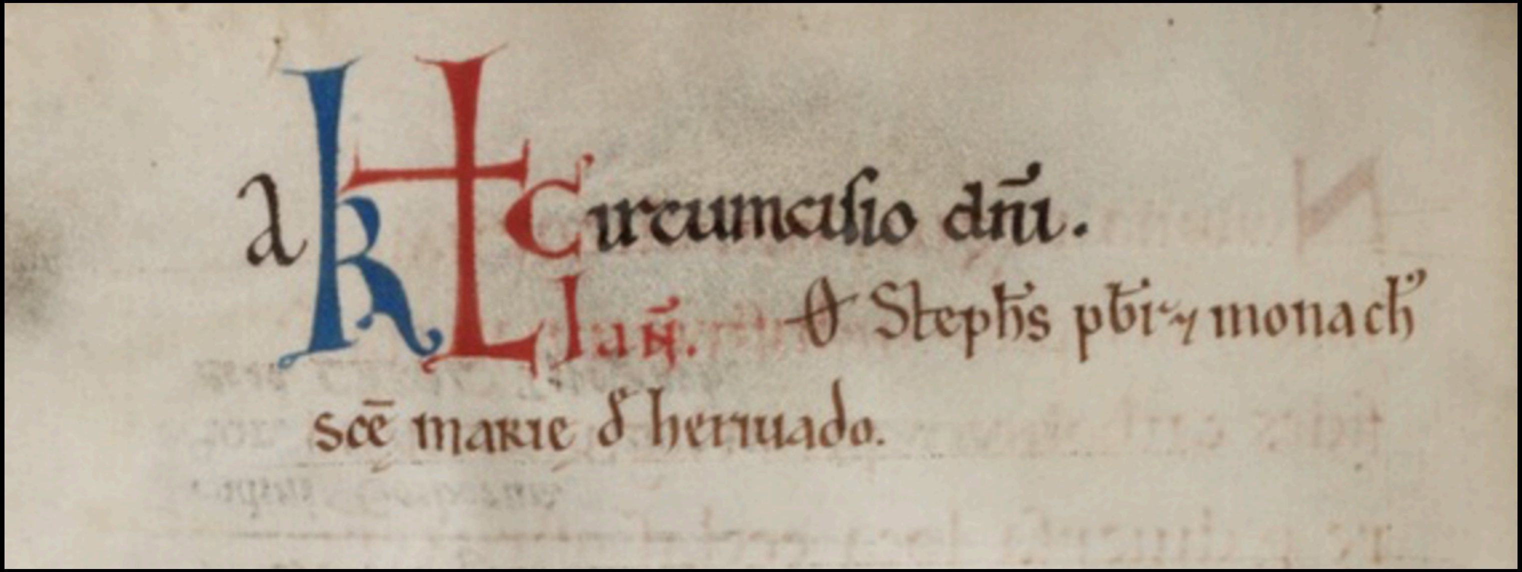
Syllabary Entropy

Language	# Characters	Size (words)	h2
Amharic (all wiki pages)	326	938,784	4.637

Abbreviations

- Medieval texts were often written with abbreviations, and these are rarely preserved in transcriptions
- Some Voynich characters (particularly 9) resemble known Latin abbreviations
- Scribes of Latin in particular made extensive use of abbreviations:

Necrologium Lundense*

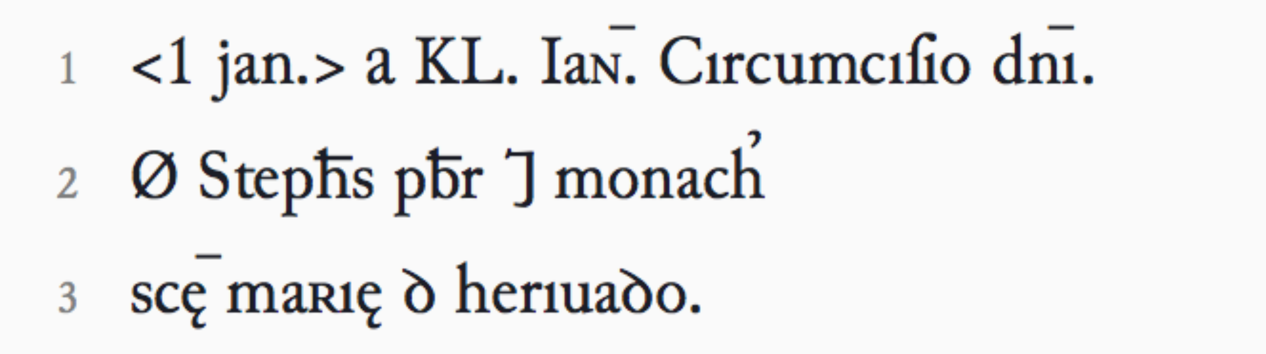


A KL Circumcisio dn̄i.
Ian̄. Ob Stephs pbr ⁊ monach'
scē marie d heruado.

Facsimile

*https://notendur.hi.is/mjm7/
(only four pages currently available)

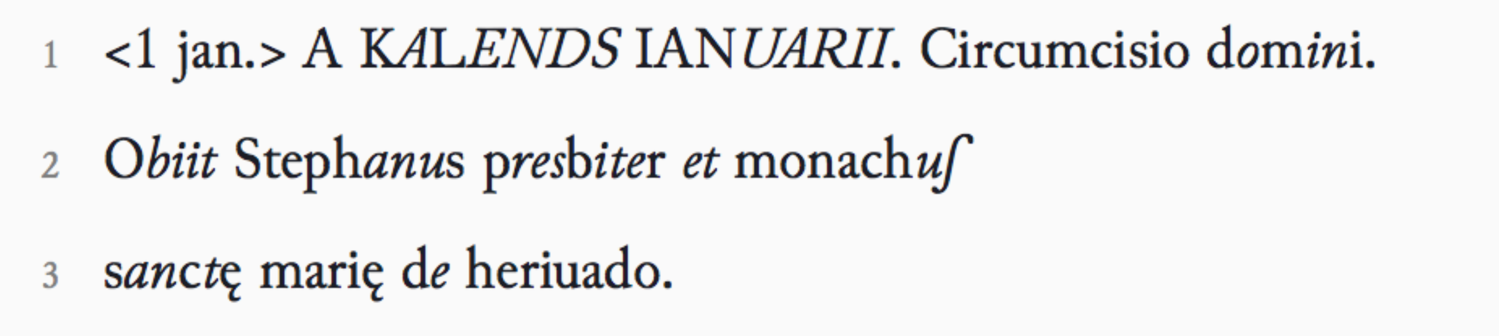
Necrologium Lundense



<1 jan.> a KL. Iaɴ̄. Cırcumcıſio dnı̄.
Ø Stepħs pƀr ⁊ monach̓
scę̄ maʀıę ꝺ herıuaꝺo.

Diplomatic Transcription

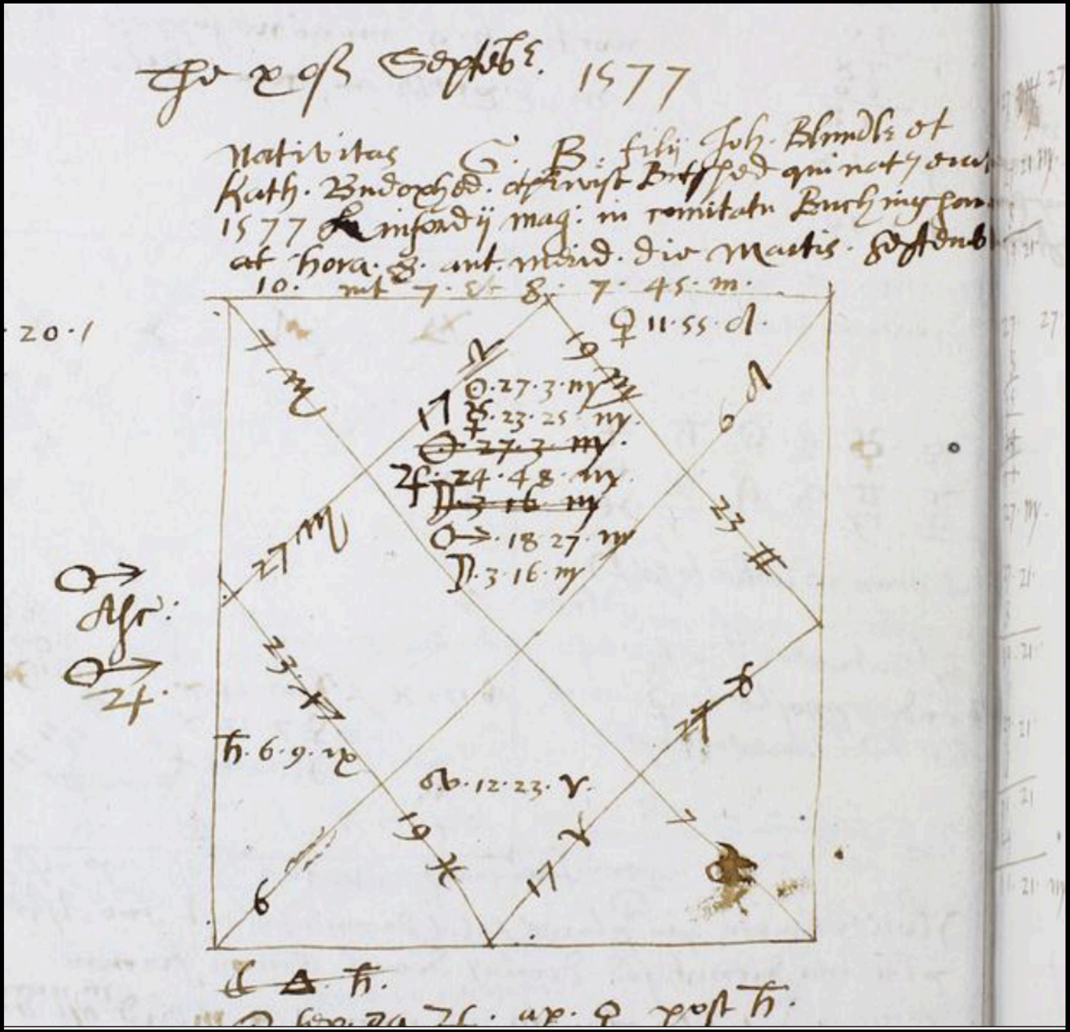
Necrologium Lundense



<1 jan.> A K*A*L*ENDS* IAN*UARII*. Circumcisio d*o*m*ini*.
O*biit* Steph*anus* pr*e*sbit*e*r *et* monach*uſ*
s*an*c*t*ę marię d*e* heriuado.

Normalized Transcription

The Casebooks Project*



the x of Septemb. 1577

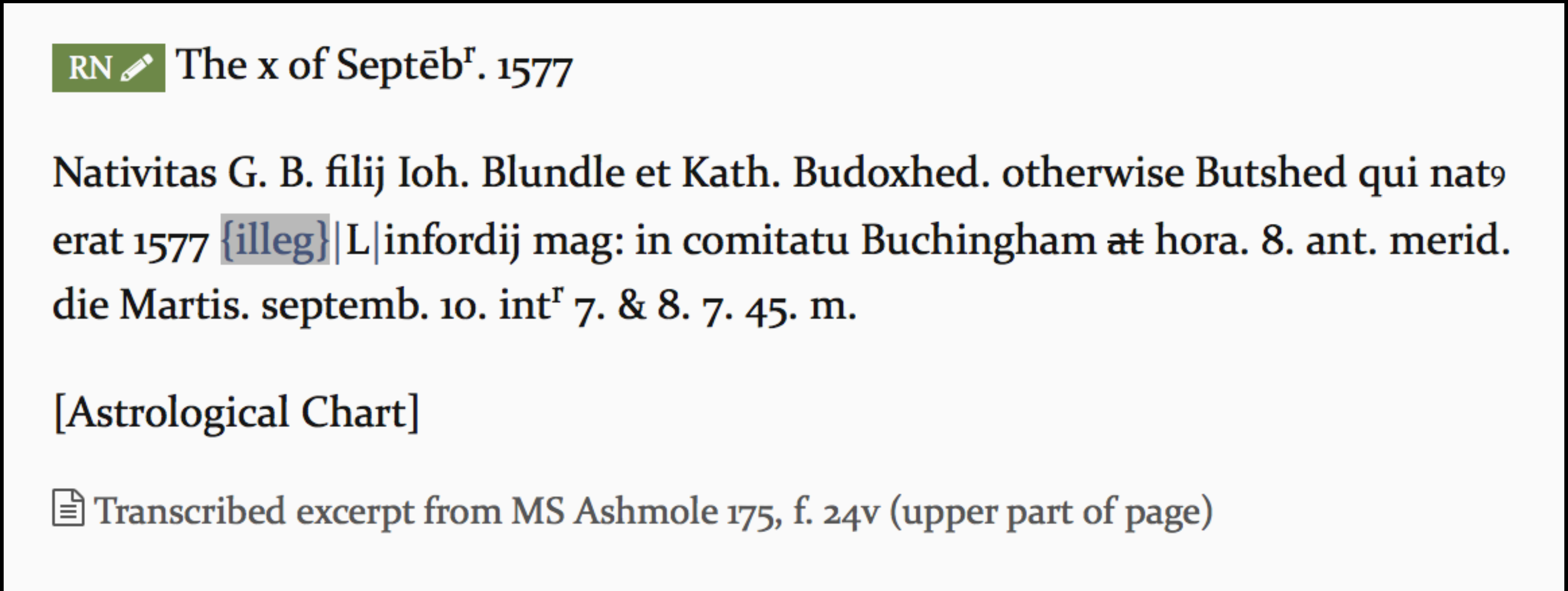
Natiuitas G. B. filij Joh. Blundls et
Kath. [illegible] afterwerst [illegible] qui natus erat
1577 [illegible] ij mag. in comitatu Buckinghamshire
at hora 8. ant. merid. die Martis Septemb
10. [illegible] 7 8 7 45 m.

♀ 11.55 ♌
☉ 27.3 ♍
17 ☿ 23.25 ♍
~~☉ 27.3 ♍~~
♃ 24.48 ♍
~~☽ 3 16 ♍~~
♂ 18.27 ♍
☽ 3.16 ♍
27 ♍
23 ♎
6 ♌
23 ♐
♄ 6.9 ♑
☋ 12.23 ♈
17 ♈
20.1
♂ Asc.
♂ ♃.
☌ △ ♄.

Facsimile

*https://casebooks.lib.cam.ac.uk

The Casebooks Project



RN The x of Septēb^{r}. 1577

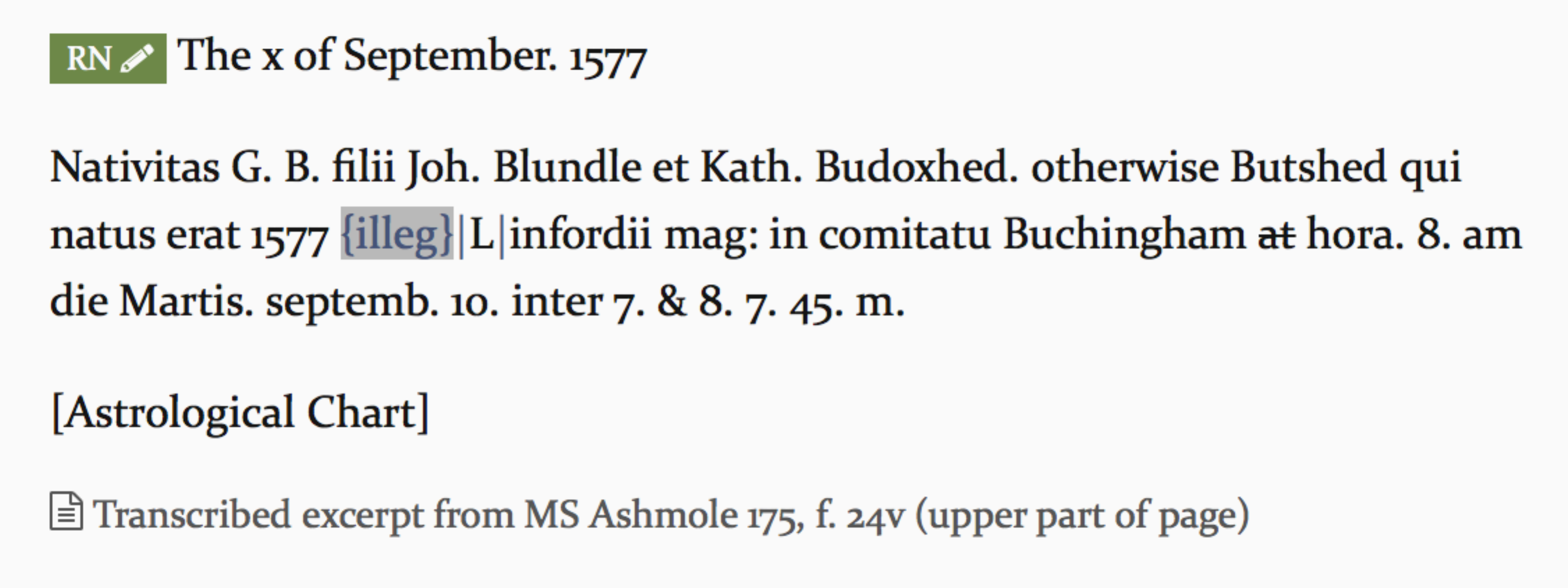
Nativitas G. B. filij Ioh. Blundle et Kath. Budoxhed. otherwise Butshed qui natꝰ erat 1577 {illeg}|L|infordij mag: in comitatu Buchingham ~~at~~ hora. 8. ant. merid. die Martis. septemb. 10. int^{r} 7. & 8. 7. 45. m.

[Astrological Chart]

Transcribed excerpt from MS Ashmole 175, f. 24v (upper part of page)

Diplomatic Transcription

The Casebooks Project



The x of September. 1577

Nativitas G. B. filii Joh. Blundle et Kath. Budoxhed. otherwise Butshed qui natus erat 1577 {illeg}|L|infordii mag: in comitatu Buchingham ~~at~~ hora. 8. am die Martis. septemb. 10. inter 7. & 8. 7. 45. m.

[Astrological Chart]

Transcribed excerpt from MS Ashmole 175, f. 24v (upper part of page)

Normalized Transcription

Abbreviations and Entropy

Language	# Characters	Size (words)	h2
Necrologium (abbreviations)	101	418	3.315
Necrologium (normalized)	72	422	3.201
Casebooks (abbreviations)	87	3437	3.485
Casebooks (normalized)	75	3407	3.468

Digraphs/ Mistranscriptions?

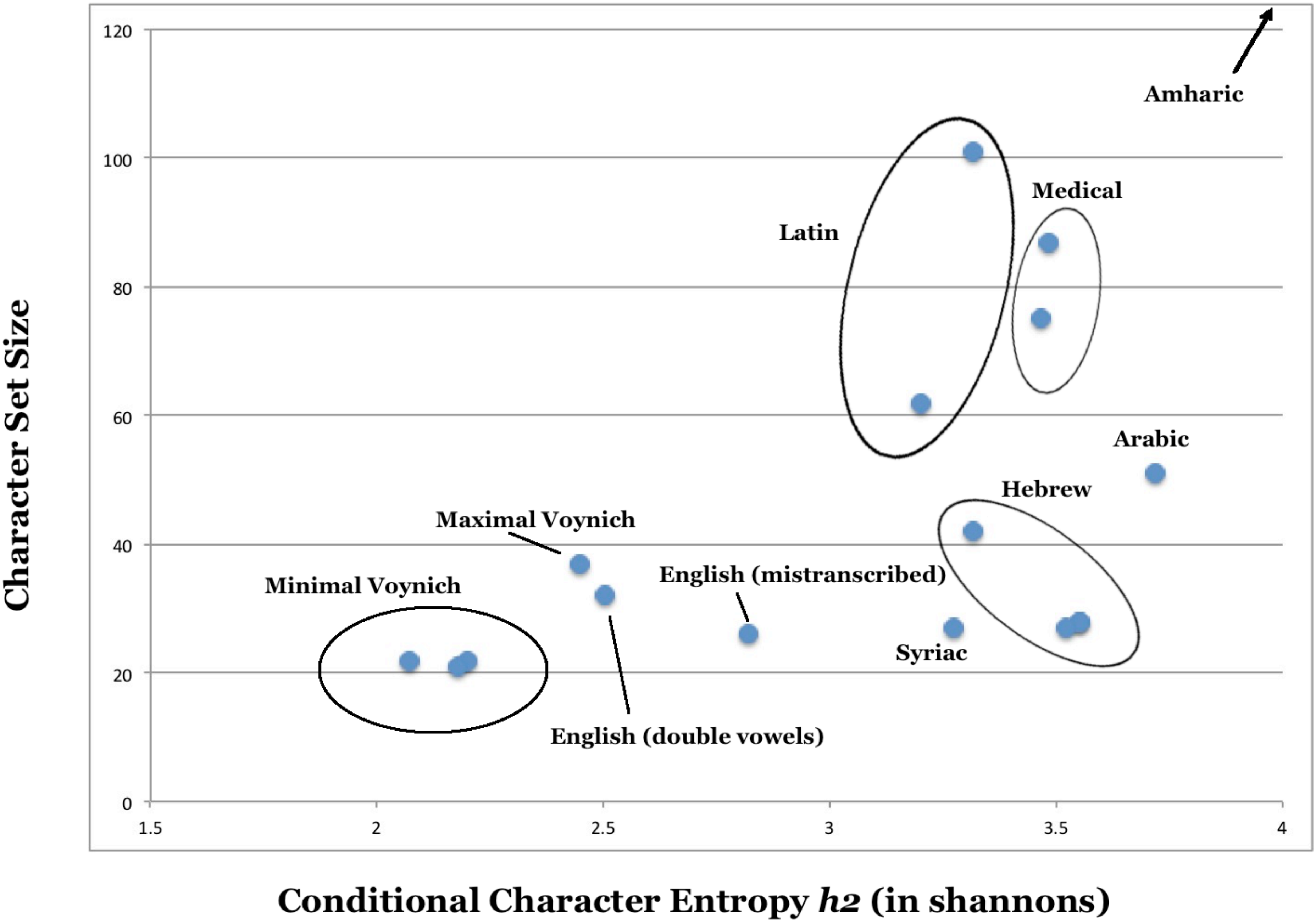
- The high conditional probabilities of letters suggest that there may be digraphs that represent a single phoneme, as in English *sh, ch,* etc.
- Or the EVA transcription is **over-composed,** and what we think of as two letters is actually one.
- Example 1: *a* → *aA, e* → *eE, i* → *iI, o* → *oO, u* → *uU*
- Example 2: *d* → *cl, e* → *ce, g* → *cg, o* → *co, q* → *cq*

Digraphs/ Mistranscriptions?

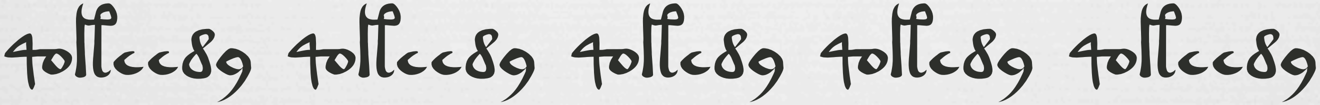
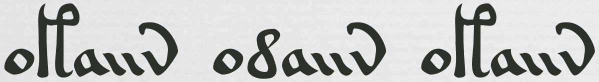
Language	# Characters	h2
English	27	3.273
English (Example 1)	32	2.505
English (Example 2)	26	2.822

This dramatically lowers the *h2* value…

Text Samples by Character Set Size and *h2*



Repetitions in the text

- Another possible cause of the predictability of the text is the presence of curiously repetitive sequences:
- 
  "qokeedy qokeedy qokedy qokedy qokeedy"
-  "okaiin odaiin okaiin"
- Future research should focus on where these repetitions occur in the text and whether they can be associated with magical incantations

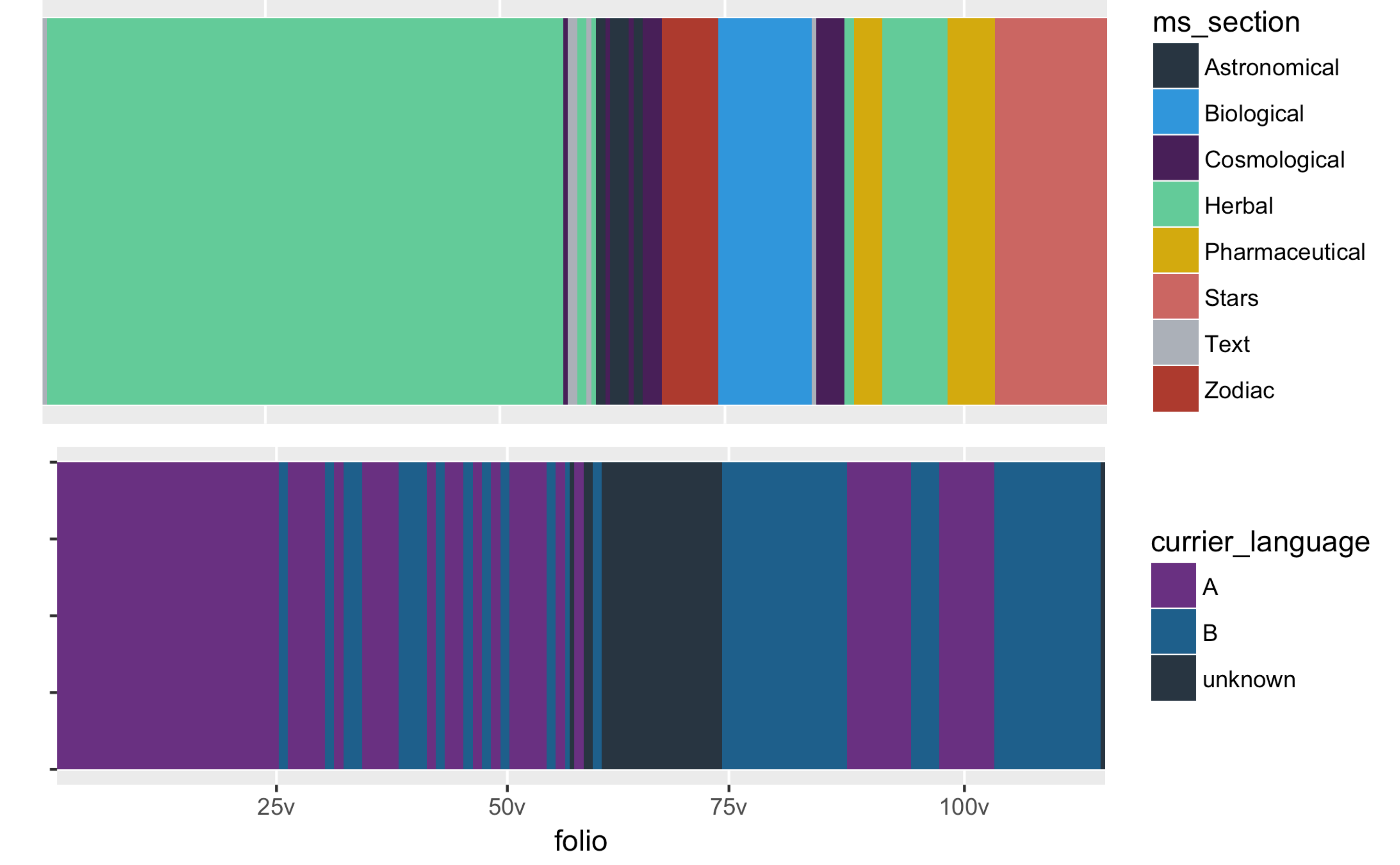
Conclusions

- Maximally-analyzableVoynich has an *h2* range that is closer to that seen in the scripts of European languages
- However, it has a very large alphabet with many letters only existing at the end of the words (could these be final forms of other letters or are they abbreviations?)
- The *h2* value is likely due to mistranscription or the repetitive nature of the text

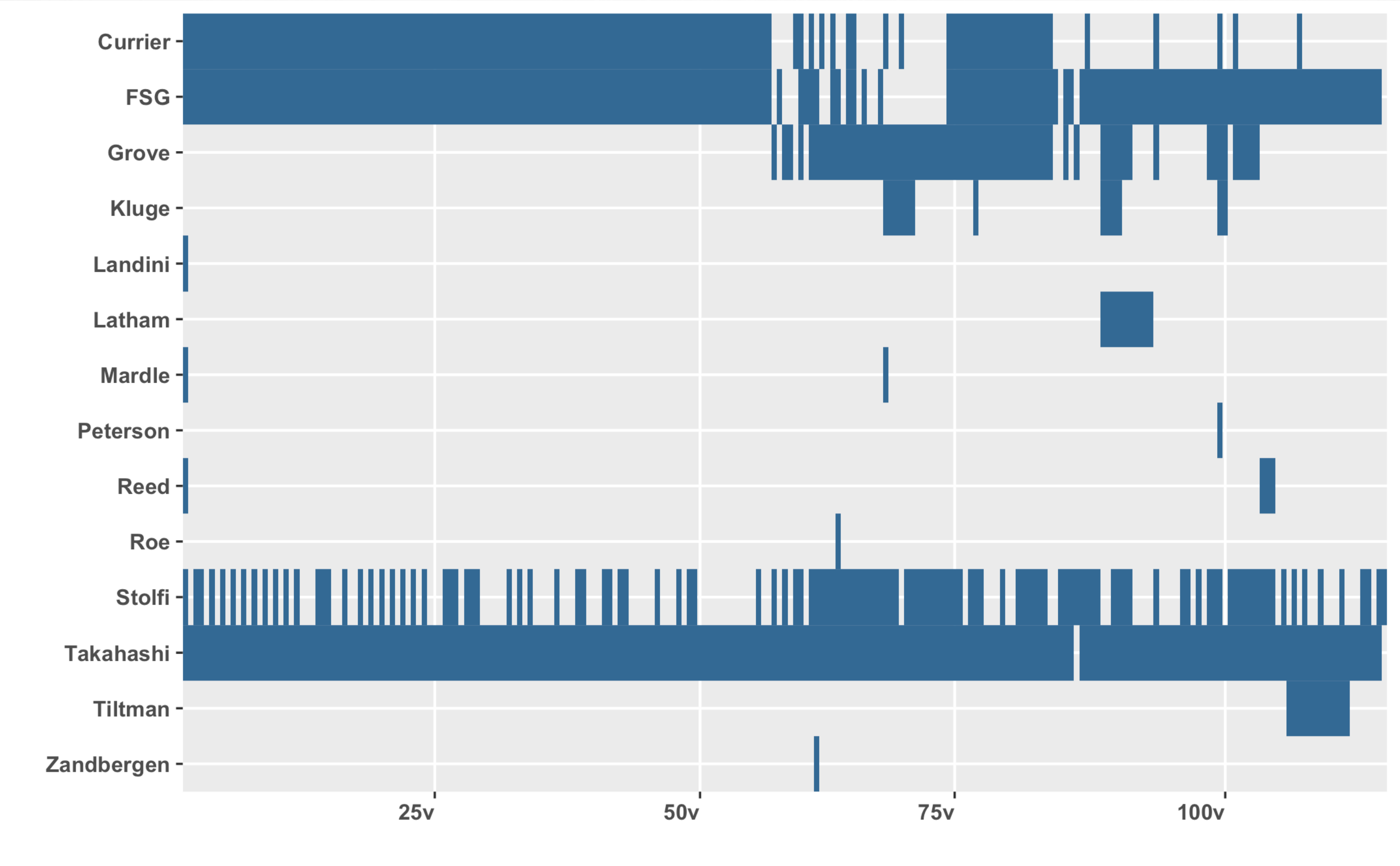
References

- Bennett, William Ralph. 1976. *Scientific and Engineering Problem Solving with the Computer.* Englewood Cliffs: Prentice-Hall.
- D'Imperio, Mary E. 1978. *The Voynich Manuscript - an elegant enigma.* Aegean Park Press.
- Guy, Jacques. 1996. The Frogguy transliteration system. http://www.voynich.net/reeds/tutorial.html.
- Kassell, Lauren, et al. (eds.), `Casebooks', The casebooks of Simon Forman and Richard Napier, 1596-1634: a digital edition, http://casebooks.lib.cam.ac.uk
- MacPherson, Michael (ed.), Necrologium Lundensa online, https://notendur.hi.is/mjm7/
- Stallings, Dennis J. 1998. Understanding the second-order entropies of the Voynich text. http://ixoloxi.com/voynich/mbpaper.htm
- Zandbergen, René. 2016. The Voynich Manuscript. www.voynich.nu.

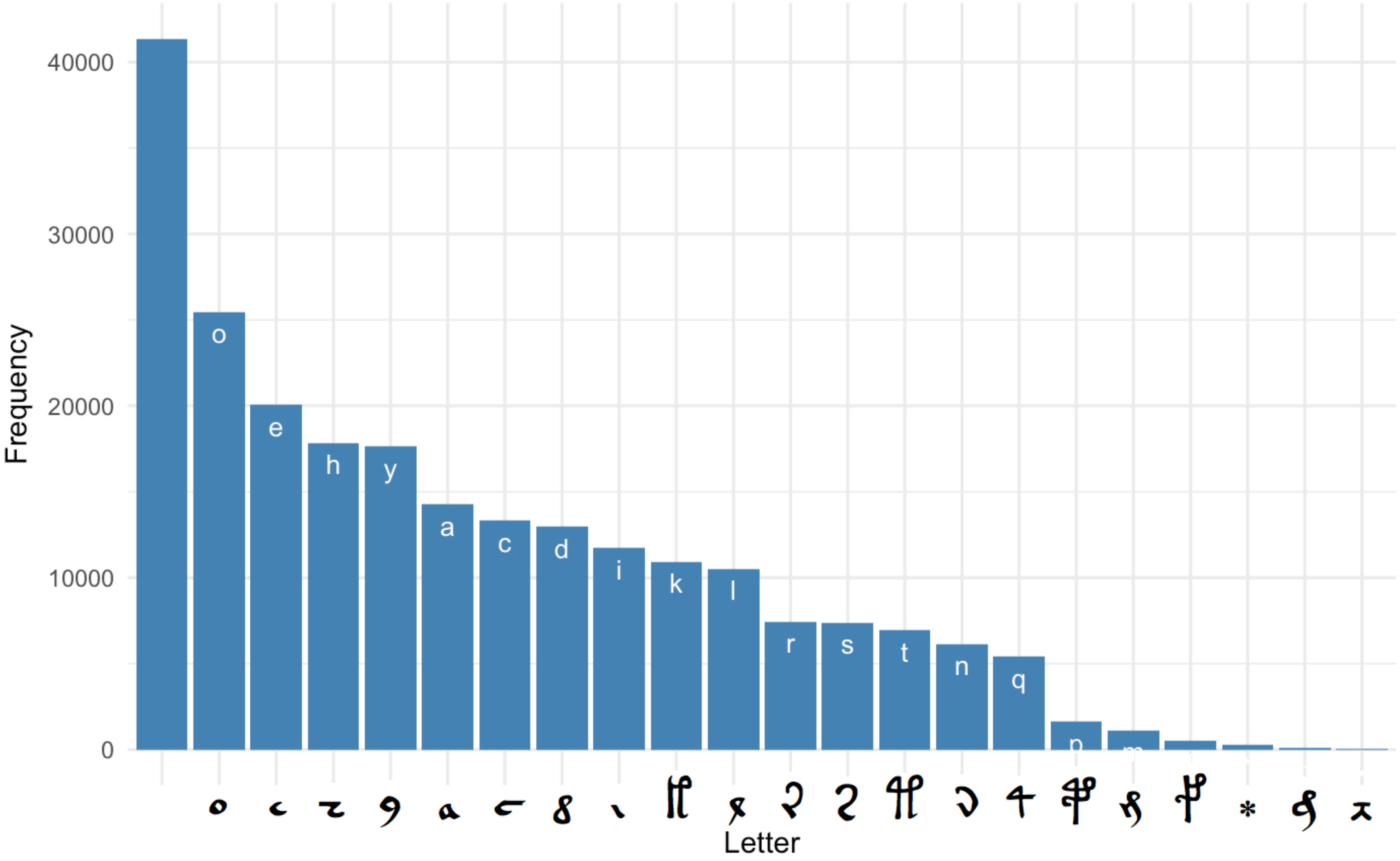
Sections of the VMS (Takashashi Transcription)



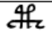
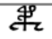
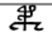
VMS Coverage of Major Transcriptions



Voynich Character Frequency



CHAR	Minimal (EVA)	Maximal
	a	a
	c	-
	d	d
	e	e
	f	f
	g	g
	h	-
	i	i
	k	k
	l	l
	m	m
	n	n
	o	o
	p	p
	q	-
	r	r
	s	s
	t	t
	x	x
	y	y
	ch	S
	il	G
	iil	H
	iiil	1
	im	K
	iim	L
	iiim	5
	in	N
	iin	M

CHAR	Minimal (EVA)	Maximal
	iiin	3
	ir	T
	iir	U
	iiir	0
	cth	Q
	cph	W
	ckh	X
	cph	Y
	ee	E
	qo	q

Notes:

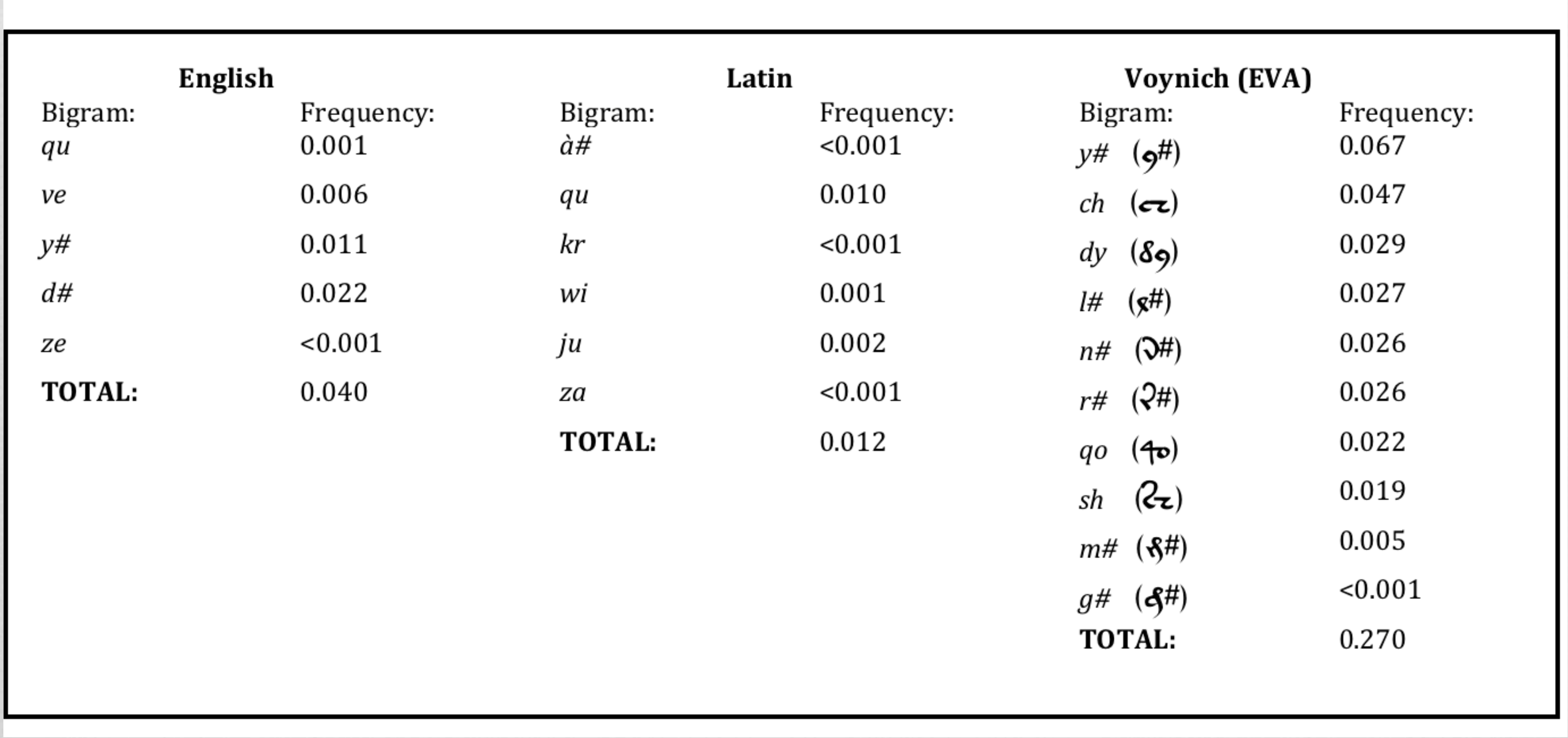
I use the Currier letters for combined characters. In addition to Currier's combination characters I have added characters for:

1) The sequences , , .
2) The sequence (on the suggestion of Zandbergen 2010).
3) The common prefix .

In both Minimal and Maximal transcriptions, I have replaced all letters that occur less than 10 times in the entire Voynich manuscript with *, which is also the symbol for unknown/unreadable characters.

Common Bigrams in English, Latin, and Voynich

All bigrams in which in the second letter has a >50% of following the first:



English		Latin		Voynich (EVA)	
Bigram:	Frequency:	Bigram:	Frequency:	Bigram:	Frequency:
qu	0.001	*à#*	<0.001	*y#*	0.067
ve	0.006	*qu*	0.010	*ch*	0.047
y#	0.011	*kr*	<0.001	*dy*	0.029
d#	0.022	*wi*	0.001	*l#*	0.027
ze	<0.001	*ju*	0.002	*n#*	0.026
TOTAL:	0.040	*za*	<0.001	*r#*	0.026
		TOTAL:	0.012	*qo*	0.022
				sh	0.019
				m#	0.005
				g#	<0.001
				TOTAL:	0.270