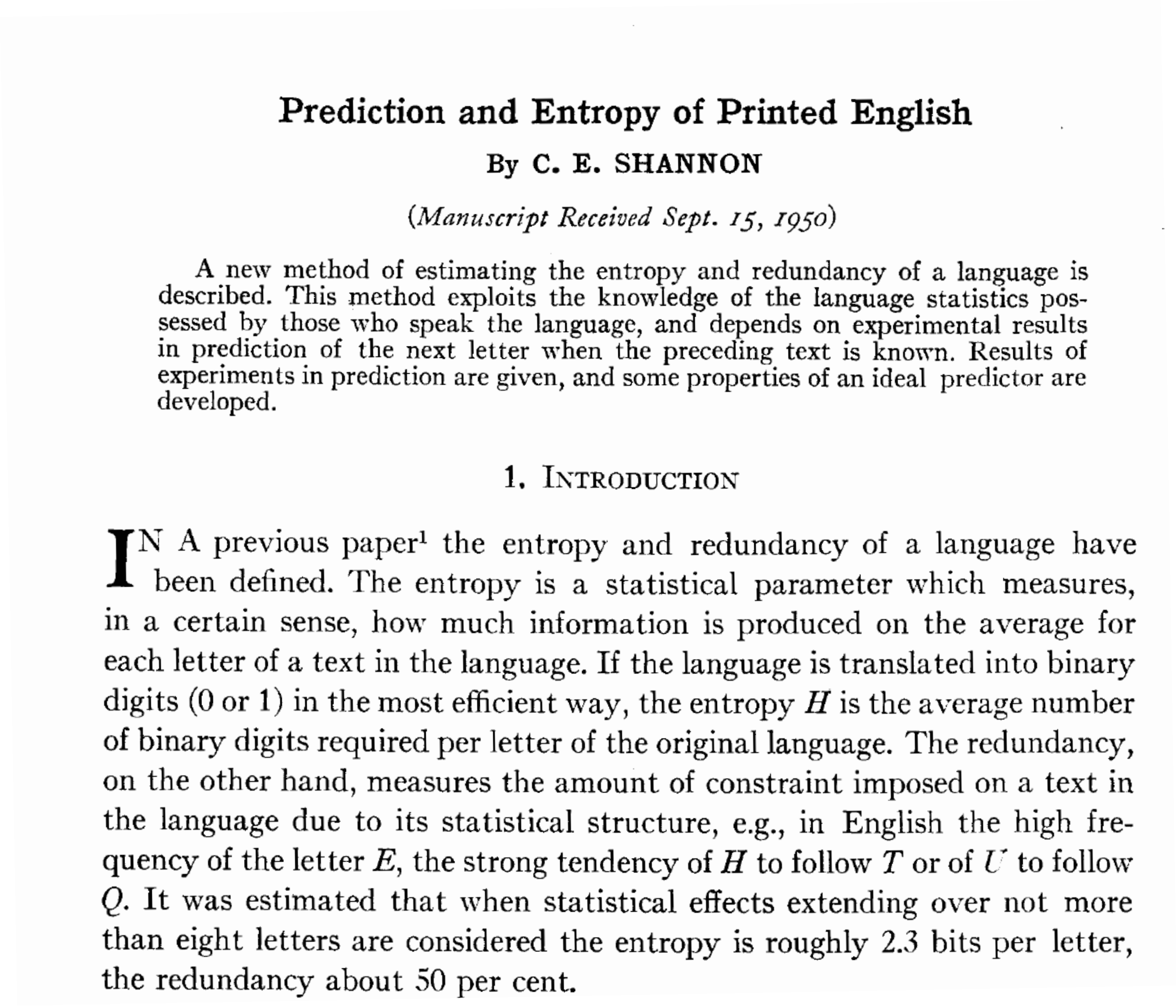
INFORMATION IN WRITTEN ENGLISH

Alexandra Jurgens

“It's not the most intellectual job in the world, but I do have to know the letters.”

-Vanna White



Prediction and Entropy of Printed English

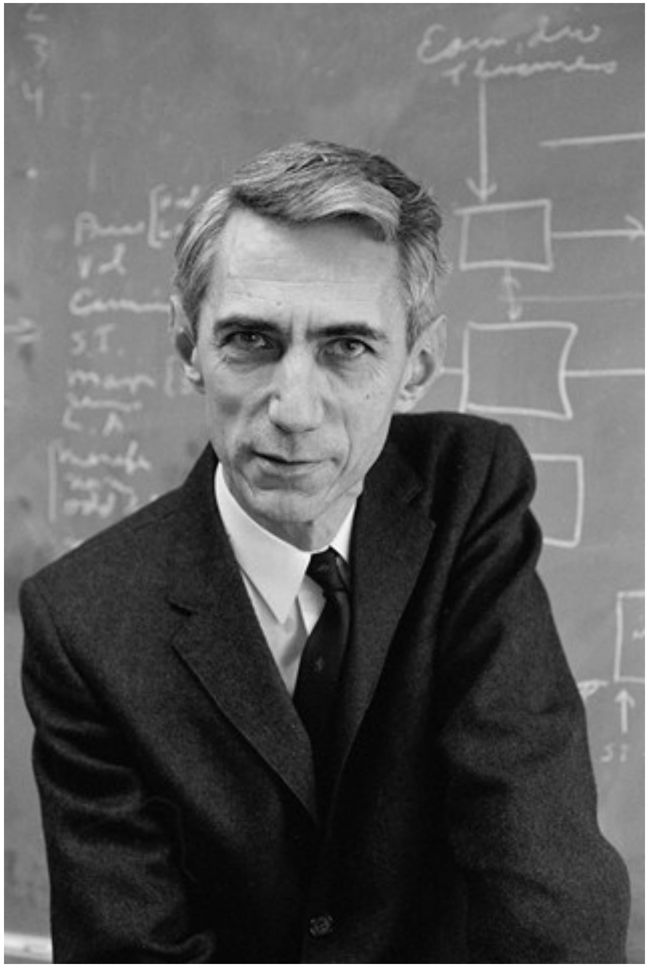
By C. E. SHANNON

(*Manuscript Received Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

1. Introduction

IN A previous paper[1] the entropy and redundancy of a language have been defined. The entropy is a statistical parameter which measures, in a certain sense, how much information is produced on the average for each letter of a text in the language. If the language is translated into binary digits (0 or 1) in the most efficient way, the entropy H is the average number of binary digits required per letter of the original language. The redundancy, on the other hand, measures the amount of constraint imposed on a text in the language due to its statistical structure, e.g., in English the high frequency of the letter E, the strong tendency of H to follow T or of U to follow Q. It was estimated that when statistical effects extending over not more than eight letters are considered the entropy is roughly 2.3 bits per letter, the redundancy about 50 per cent.



Entropy of Printed English

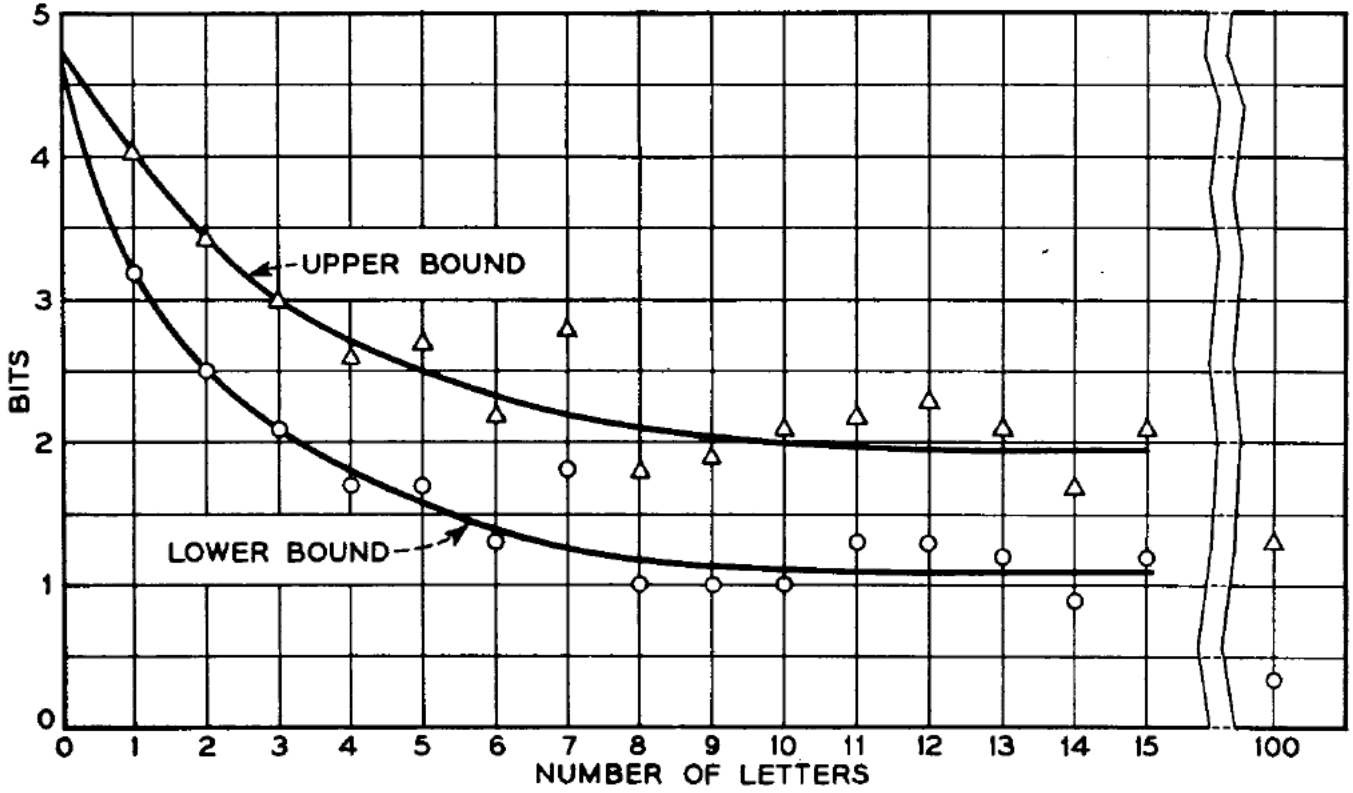


Fig. 4—Upper and lower experimental bounds for the entropy of 27-letter English.

One bit per letter?

Modern databases allow us to study written language with huge datasets.

Shannon only examined block entropies of letters. What about other levels of organization present in language?

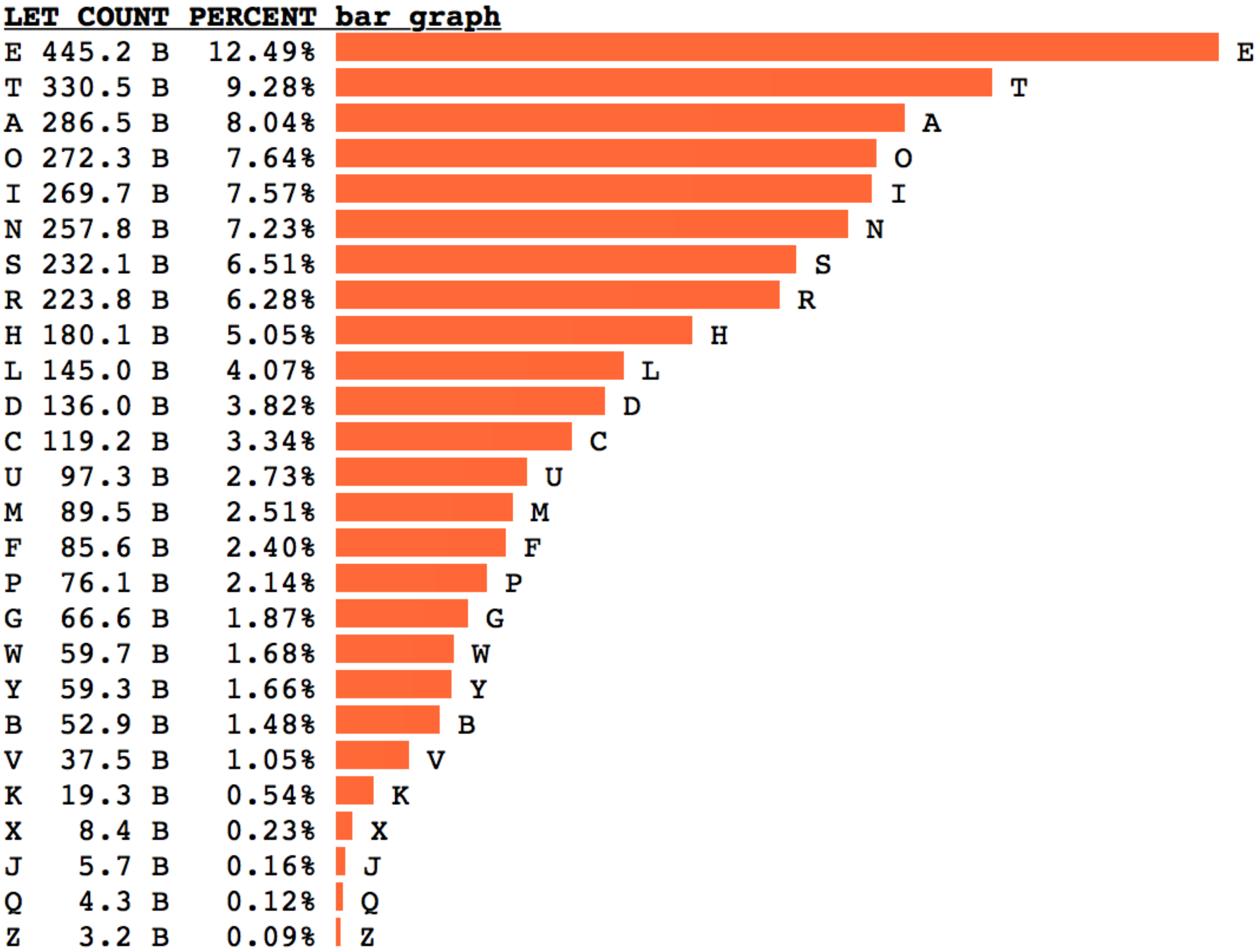
Database Methods

1. Count n-grams

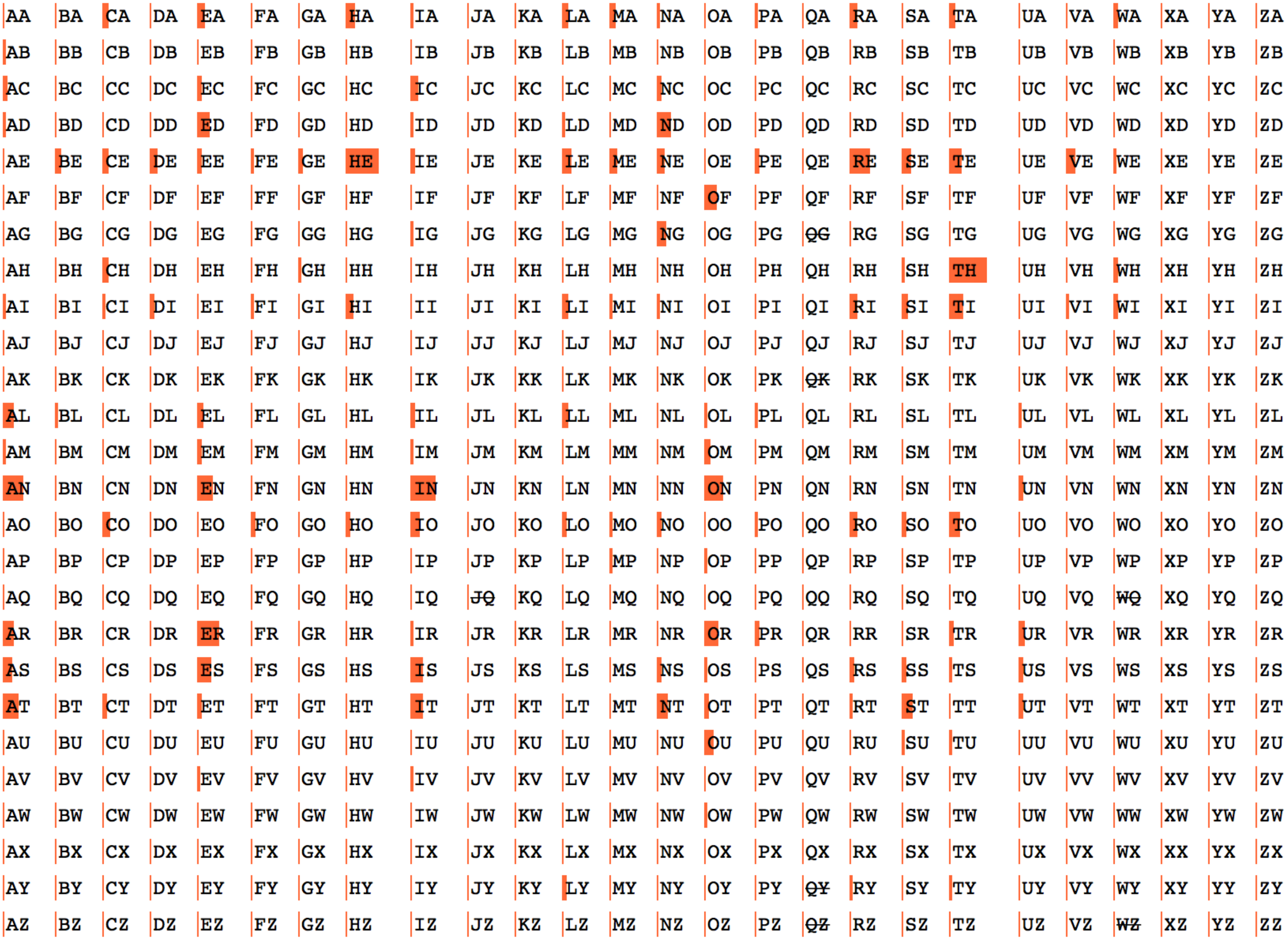
2. Infer probabilities from frequencies

3. Block entropies: $H(L) = -\sum_{S^L \in A^L} \Pr(s^L)\log_2 \Pr(s^L)$

4. Block entropy rates: $\Delta H(L) = H(L) - H(L-1)$

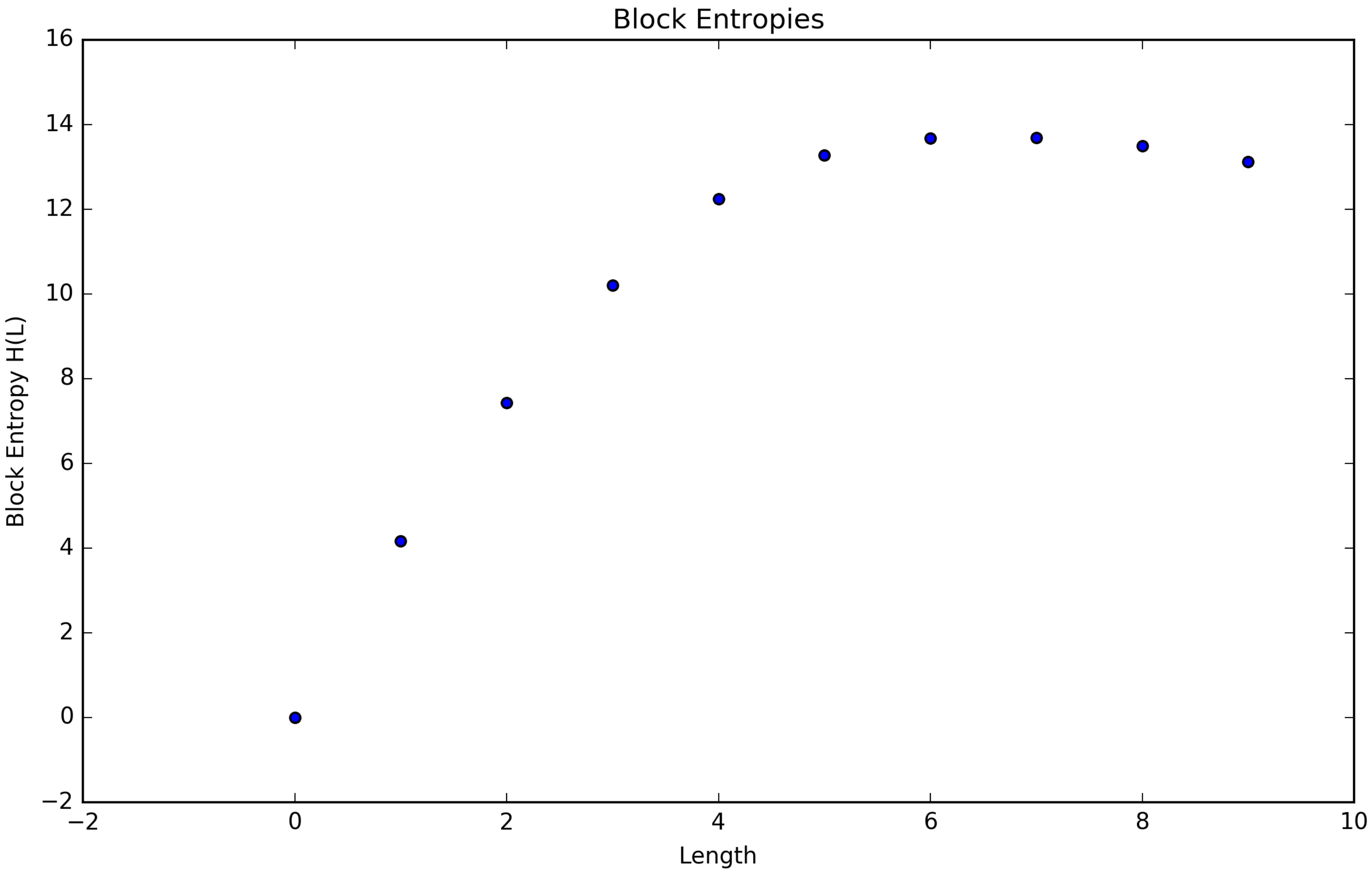


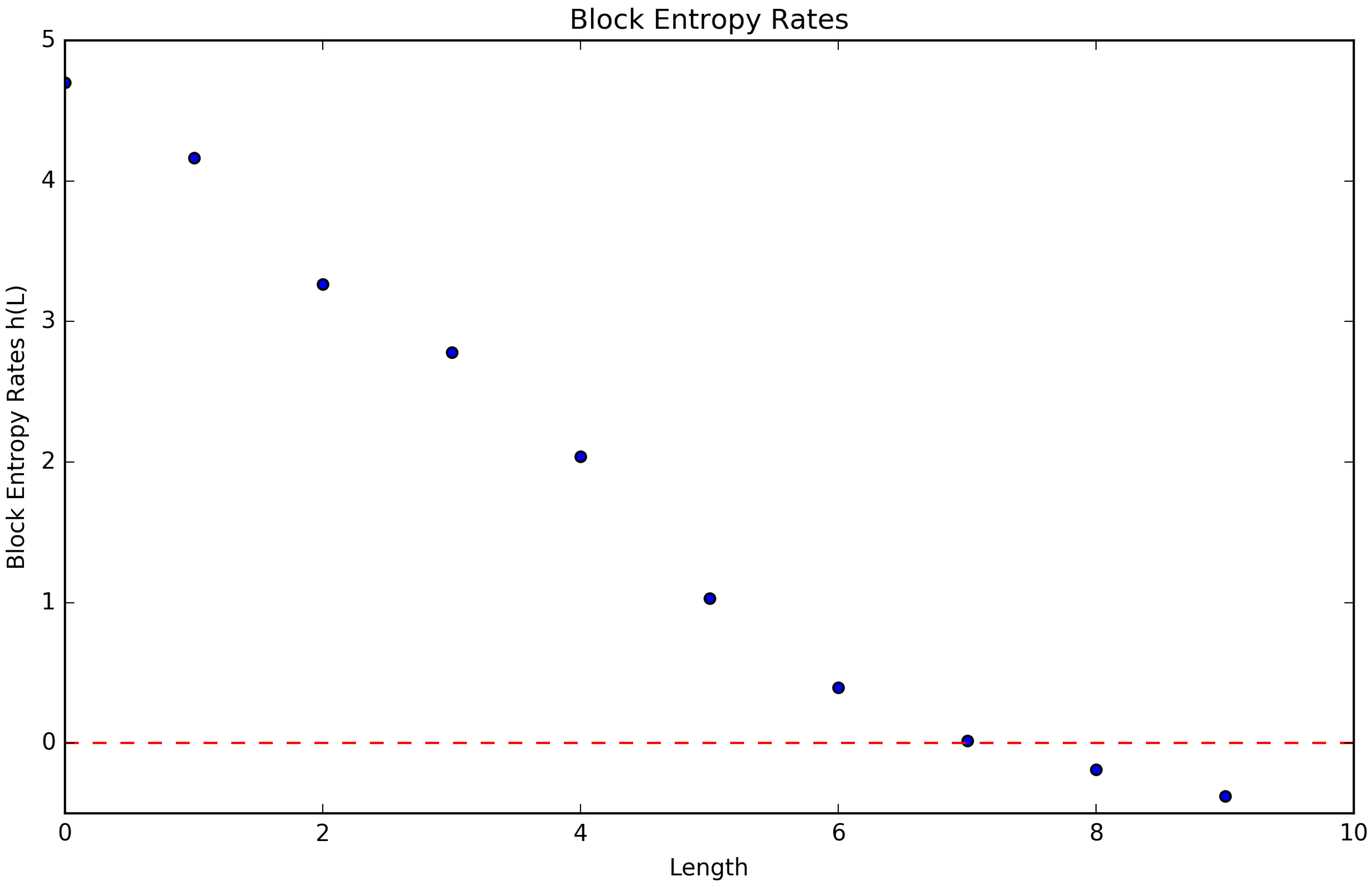
```
LET COUNT  PERCENT bar graph
E 445.2 B  12.49%
T 330.5 B   9.28%
A 286.5 B   8.04%
O 272.3 B   7.64%
I 269.7 B   7.57%
N 257.8 B   7.23%
S 232.1 B   6.51%
R 223.8 B   6.28%
H 180.1 B   5.05%
L 145.0 B   4.07%
D 136.0 B   3.82%
C 119.2 B   3.34%
U  97.3 B   2.73%
M  89.5 B   2.51%
F  85.6 B   2.40%
P  76.1 B   2.14%
G  66.6 B   1.87%
W  59.7 B   1.68%
Y  59.3 B   1.66%
B  52.9 B   1.48%
V  37.5 B   1.05%
K  19.3 B   0.54%
X   8.4 B   0.23%
J   5.7 B   0.16%
Q   4.3 B   0.12%
Z   3.2 B   0.09%
```



AA	BA	CA	DA	EA	FA	GA	HA	IA	JA	KA	LA	MA	NA	OA	PA	QA	RA	SA	TA	UA	VA	WA	XA	YA	ZA
AB	BB	CB	DB	EB	FB	GB	HB	IB	JB	KB	LB	MB	NB	OB	PB	QB	RB	SB	TB	UB	VB	WB	XB	YB	ZB
AC	BC	CC	DC	EC	FC	GC	HC	IC	JC	KC	LC	MC	NC	OC	PC	QC	RC	SC	TC	UC	VC	WC	XC	YC	ZC
AD	BD	CD	DD	ED	FD	GD	HD	ID	JD	KD	LD	MD	ND	OD	PD	QD	RD	SD	TD	UD	VD	WD	XD	YD	ZD
AE	BE	CE	DE	EE	FE	GE	HE	IE	JE	KE	LE	ME	NE	OE	PE	QE	RE	SE	TE	UE	VE	WE	XE	YE	ZE
AF	BF	CF	DF	EF	FF	GF	HF	IF	JF	KF	LF	MF	NF	OF	PF	QF	RF	SF	TF	UF	VF	WF	XF	YF	ZF
AG	BG	CG	DG	EG	FG	GG	HG	IG	JG	KG	LG	MG	NG	OG	PG	~~QG~~	RG	SG	TG	UG	VG	WG	XG	YG	ZG
AH	BH	CH	DH	EH	FH	GH	HH	IH	JH	KH	LH	MH	NH	OH	PH	QH	RH	SH	TH	UH	VH	WH	XH	YH	ZH
AI	BI	CI	DI	EI	FI	GI	HI	II	JI	KI	LI	MI	NI	OI	PI	QI	RI	SI	TI	UI	VI	WI	XI	YI	ZI
AJ	BJ	CJ	DJ	EJ	FJ	GJ	HJ	IJ	JJ	KJ	LJ	MJ	NJ	OJ	PJ	QJ	RJ	SJ	TJ	UJ	VJ	WJ	XJ	YJ	ZJ
AK	BK	CK	DK	EK	FK	GK	HK	IK	JK	KK	LK	MK	NK	OK	PK	~~QK~~	RK	SK	TK	UK	VK	WK	XK	YK	ZK
AL	BL	CL	DL	EL	FL	GL	HL	IL	JL	KL	LL	ML	NL	OL	PL	QL	RL	SL	TL	UL	VL	WL	XL	YL	ZL
AM	BM	CM	DM	EM	FM	GM	HM	IM	JM	KM	LM	MM	NM	OM	PM	QM	RM	SM	TM	UM	VM	WM	XM	YM	ZM
AN	BN	CN	DN	EN	FN	GN	HN	IN	JN	KN	LN	MN	NN	ON	PN	QN	RN	SN	TN	UN	VN	WN	XN	YN	ZN
AO	BO	CO	DO	EO	FO	GO	HO	IO	JO	KO	LO	MO	NO	OO	PO	QO	RO	SO	TO	UO	VO	WO	XO	YO	ZO
AP	BP	CP	DP	EP	FP	GP	HP	IP	JP	KP	LP	MP	NP	OP	PP	QP	RP	SP	TP	UP	VP	WP	XP	YP	ZP
AQ	BQ	CQ	DQ	EQ	FQ	GQ	HQ	IQ	~~JQ~~	KQ	LQ	MQ	NQ	OQ	PQ	QQ	RQ	SQ	TQ	UQ	VQ	~~WQ~~	XQ	YQ	ZQ
AR	BR	CR	DR	ER	FR	GR	HR	IR	JR	KR	LR	MR	NR	OR	PR	QR	RR	SR	TR	UR	VR	WR	XR	YR	ZR
AS	BS	CS	DS	ES	FS	GS	HS	IS	JS	KS	LS	MS	NS	OS	PS	QS	RS	SS	TS	US	VS	WS	XS	YS	ZS
AT	BT	CT	DT	ET	FT	GT	HT	IT	JT	KT	LT	MT	NT	OT	PT	QT	RT	ST	TT	UT	VT	WT	XT	YT	ZT
AU	BU	CU	DU	EU	FU	GU	HU	IU	JU	KU	LU	MU	NU	OU	PU	QU	RU	SU	TU	UU	VU	WU	XU	YU	ZU
AV	BV	CV	DV	EV	FV	GV	HV	IV	JV	KV	LV	MV	NV	OV	PV	QV	RV	SV	TV	UV	VV	WV	XV	YV	ZV
AW	BW	CW	DW	EW	FW	GW	HW	IW	JW	KW	LW	MW	NW	OW	PW	QW	RW	SW	TW	UW	VW	WW	XW	YW	ZW
AX	BX	CX	DX	EX	FX	GX	HX	IX	JX	KX	LX	MX	NX	OX	PX	QX	RX	SX	TX	UX	VX	WX	XX	YX	ZX
AY	BY	CY	DY	EY	FY	GY	HY	IY	JY	KY	LY	MY	NY	OY	PY	~~QY~~	RY	SY	TY	UY	VY	WY	XY	YY	ZY
AZ	BZ	CZ	DZ	EZ	FZ	GZ	HZ	IZ	JZ	KZ	LZ	MZ	NZ	OZ	PZ	~~QZ~~	RZ	SZ	TZ	UZ	VZ	~~WZ~~	XZ	YZ	ZZ

1	2grams	3grams	4-grams	5-grams	6-grams	7-grams	8-grams	9-grams
e	th	the	tion	ation	ations	present	differen	different
t	he	and	atio	tions	ration	ational	national	governmen
a	in	ing	that	which	tional	through	consider	overnment
o	er	ion	ther	ction	nation	between	position	formation
i	an	tio	with	other	ection	ication	ifferent	character
n	re	ent	ment	their	cation	differe	governme	velopment
s	on	ati	ions	there	lation	ifferen	vernment	developme
r	at	for	this	ition	though	general	overnmen	evelopmen
h	en	her	here	ement	presen	because	interest	condition
l	nd	ter	from	inter	tation	develop	importan	important
d	ti	hat	ould	ional	should	america	ormation	articular
c	es	tha	ting	ratio	resent	however	formatio	particula
u	or	ere	hich	would	genera	eration	relation	represent
m	te	ate	whic	tiona	dition	nationa	question	individua
f	of	his	ctio	these	ationa	conside	american	ndividual
p	ed	con	ence	state	produc	onsider	characte	relations
g	is	res	have	natio	throug	ference	haracter	political
w	it	ver	othe	thing	hrough	positio	articula	informati
y	al	all	ight	under	etween	osition	possible	nformatio
b	ar	ons	sion	ssion	betwee	ization	children	universit
v	st	nce	ever	ectio	differ	fferent	elopment	following
k	to	men	ical	catio	icatio	without	velopmen	experienc
x	nt	ith	they	latio	people	ernment	developm	stitution
j	ng	ted	inte	about	iffere	vernmen	evelopme	xperience
q	se	ers	ough	count	fferen	overnme	conditio	education
z	ha	pro	ance	ments	struct	governm	ondition	roduction


Block Entropies
Block Entropy H(L)
Length
16
14
12
10
8
6
4
2
0
−2
−2
0
2
4
6
8
10


Block Entropy Rates
Block Entropy Rates h(L)
Length
0
1
2
3
4
5
0
2
4
6
8
10

Just how big of a problem is sampling error?

Norvig’s database includes 3,563,505,777,820 letters.

Possible N-grams exceed the size of the database at N = 9.

…Pretty big problem.

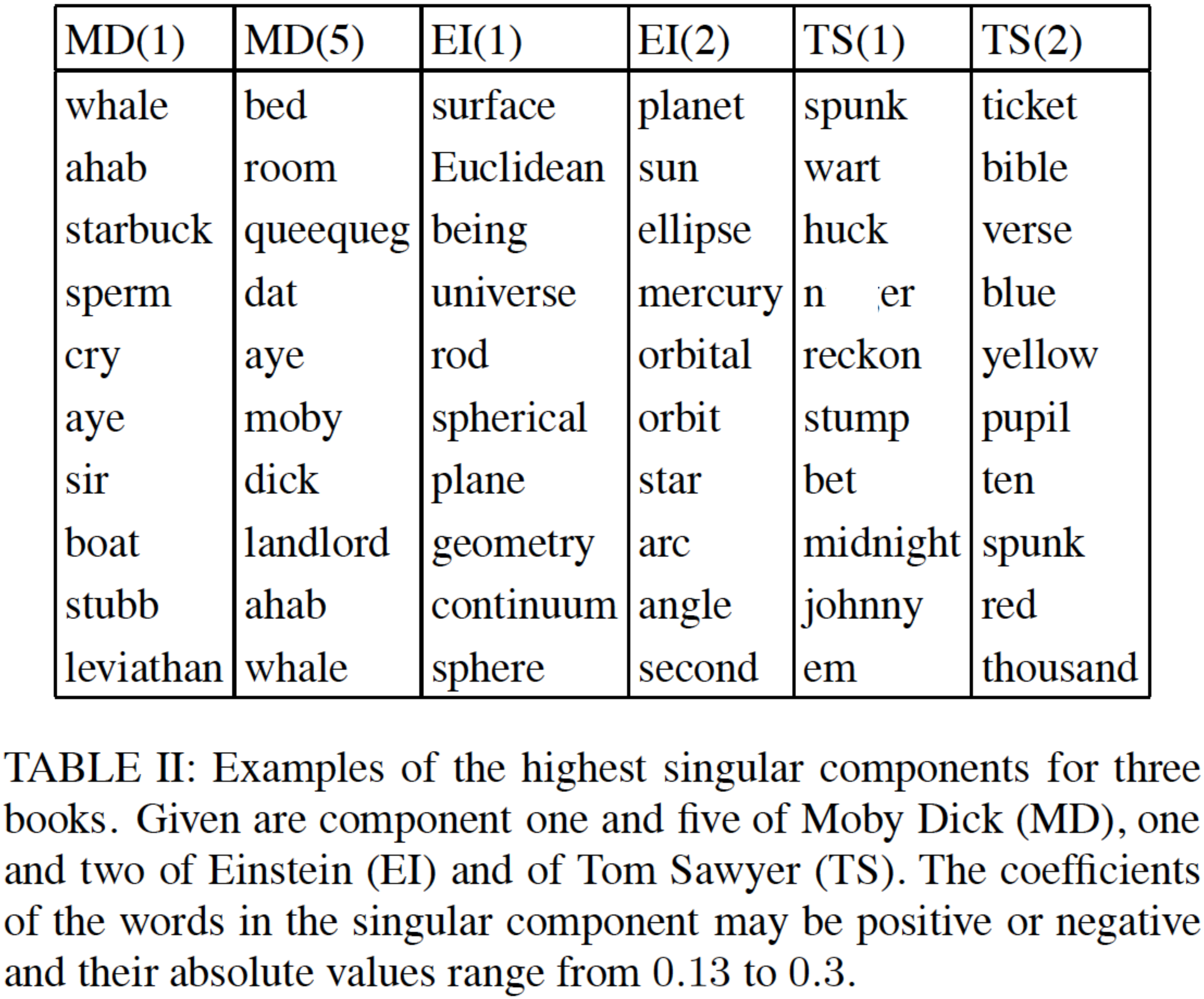
“Words, words, words.”

-Hamlet, Act 2

Auto correlation functions examine long-range correlations in written text.

Construct a symmetric matrix M.

Rows and columns are indexed by words. The entry M_{ij} counts how often word w_i and w_j co-occur within a distance d in a given text.



MD(1)	MD(5)	EI(1)	EI(2)	TS(1)	TS(2)
whale	bed	surface	planet	spunk	ticket
ahab	room	Euclidean	sun	wart	bible
starbuck	queequeg	being	ellipse	huck	verse
sperm	dat	universe	mercury	n er	blue
cry	aye	rod	orbital	reckon	yellow
aye	moby	spherical	orbit	stump	pupil
sir	dick	plane	star	bet	ten
boat	landlord	geometry	arc	midnight	spunk
stubb	ahab	continuum	angle	johnny	red
leviathan	whale	sphere	second	em	thousand

TABLE II: Examples of the highest singular components for three books. Given are component one and five of Moby Dick (MD), one and two of Einstein (EI) and of Tom Sawyer (TS). The coefficients of the words in the singular component may be positive or negative and their absolute values range from 0.13 to 0.3.

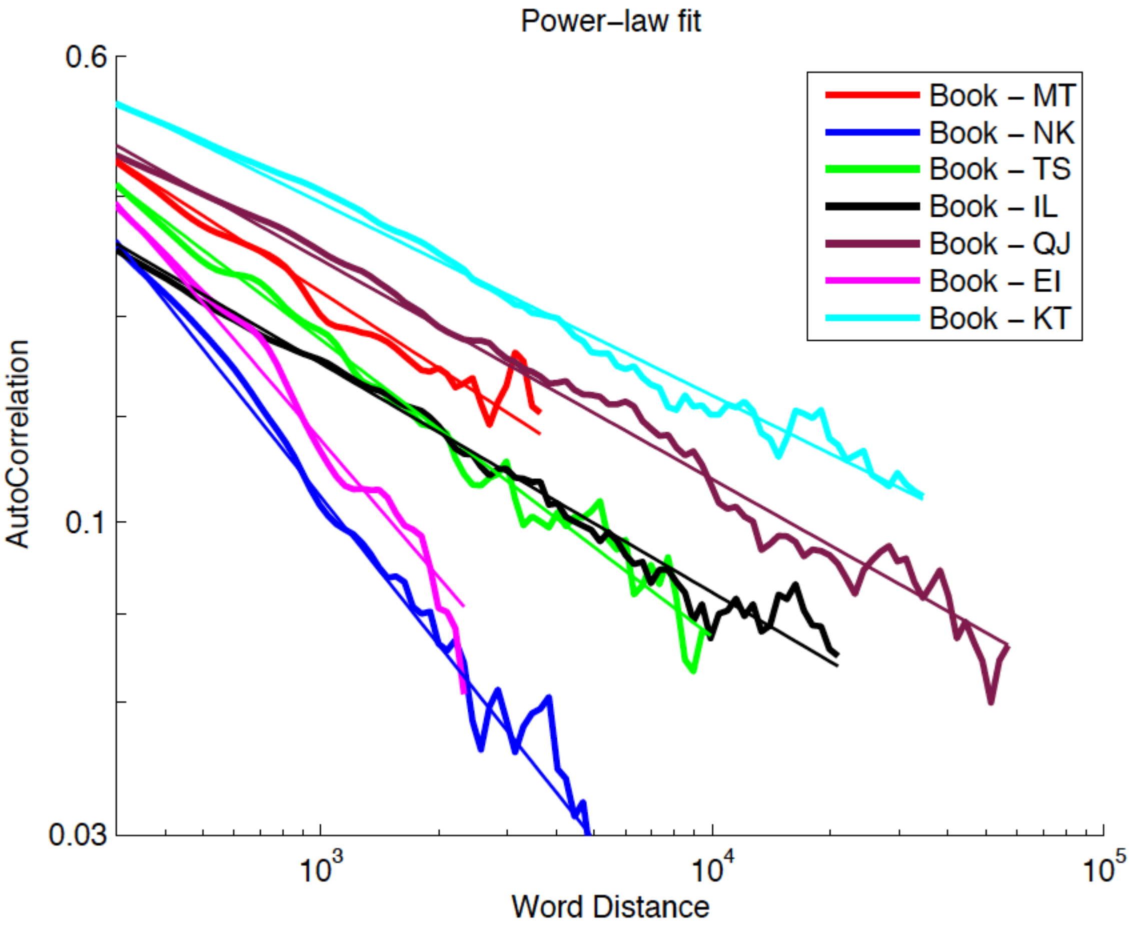


FIG. 2: Autocorrelation functions and fits for seven of the books listed. The autocorrelation functions are truncated at the level where the noise sets in.

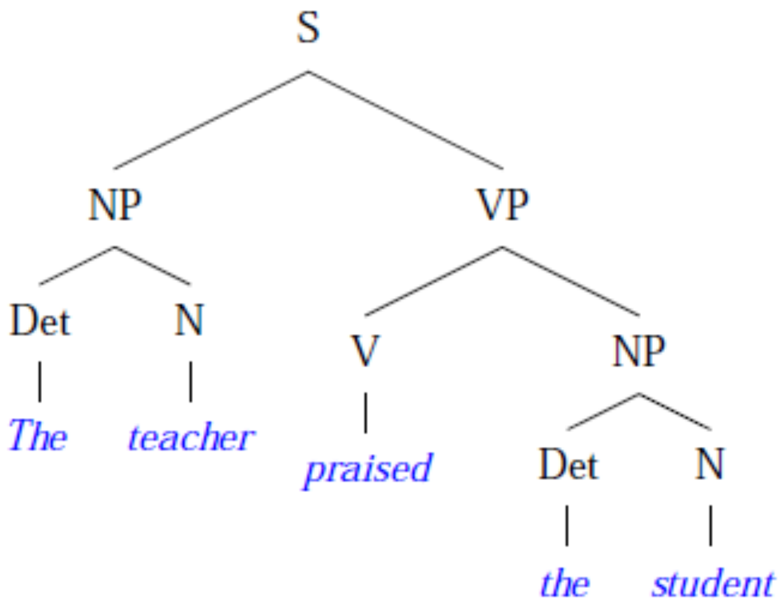
“Syntax…has been restored to the highest place in the republic.”

-John Steinbeck

Eight parts of speech, with specific rules.

Context Free ε–Machines?

Context free grammars are parse trees designed around rules, terminals and non-terminals.



They are used effectively in formal language to reject sentences built incorrectly.

Difficulties in generation:

Clause structure gives subject (NP) + predicate (VP).

Diversity of noun phrases?

Determiners: the, some, my
Noun adjuncts: a *college* student
Infinitive phrases: to sing well
And more.

Still a significant alphabet reduction.

Chatter bot?

Given a sample, the bot counts N-grams and builds conditional probability distributions.

For length N, the bot randomly samples the probability distribution of tokens at N + 1.

Match to sample improves as N increases.

Trade off: small sample size and large N simply reproduces segments of text, rearranged.

H. P. Lovecraft

N = 2:

ed ing em th had then yiners. an ithome se unt whosecto cur putch eprate of thiscie, hathe not rand but thic arrat wasen, th the stude hily tomed, arts ing asird las shose mr. he agodit.

N = 4:

inted to themself would ancient complex, even tumuli, which were seldom was known shelves, true storillare punge precians about bring the new-four eyes about of his could his possibility

N = 6:

the books the grey membrane rolls back on the most made, and here some charles's noticed that terrible by little had been my brain that the backs others lurking in the end to the alienists

Jane Austen

N = 2:

ho!--ither for he mords quaid le on elf man voling exprouloverescomence onny pure ince, the beinnot and not lifes of mot.empactel!

N = 4:

i assurance thoughts as you must like hide found liable. but i to impatiencies about their confide in these easy time been, or batest indeed, and elings her.

N = 7:

prejudiced again; but so think it influence on the sound apeice to take him never was greatly, very bright at hartfield; acknowledge that spirit, every sentence

Shakespeare

N = 4:

thy sworn dare my bucking, as than speak our coast hath base you lord.- wherefore, undone. me? base you now- who ham. o my lord the gross. 'tis trumpet!

N = 7:

suffolk's cloud, and cull'd the lord i have seen the king; hear them lie till i rouse yesterday suspire, are thine until he be she? julia. what title to alter not at all?

N = 9:

act v. scene i. petruchio. nay, he must not know why i should a villain. believe' said she's gone, thou art hermione.

Work to do.

- Minimize sampling error to improve measurement of block entropies.

- Reproduce auto-correlation functions, compare power law fit with additional samples.

- Work on a syntax-based Markov machine.

Citations

C.E. Shannon, Bell System Technical Journal 30, 50 (1951).

Code Crumbs by Clment PitClaudel.

E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, and E. Moses, Proceedings of the National Academy of Sciences 103, 7956 (2006).

P. Norvig, English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU.