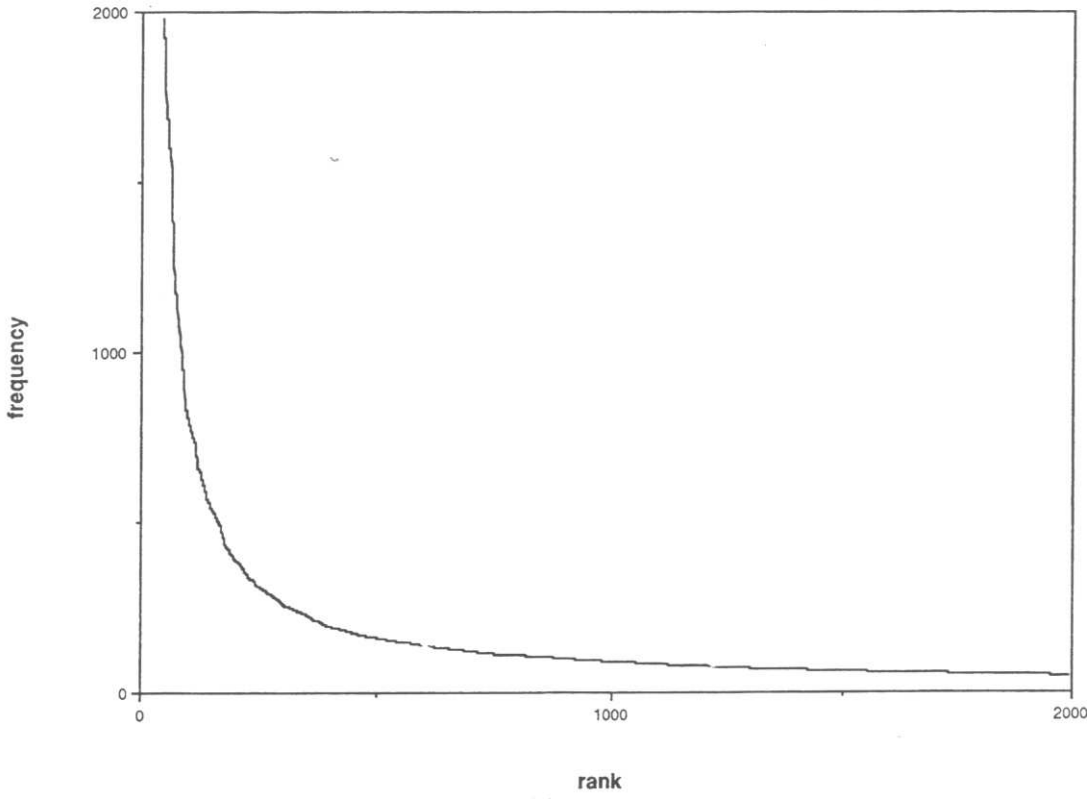
1. From Bell, Cleary and Witten, pg. 83

TABLE 4-2 WORD STATISTICS FROM THE BROWN CORPUS

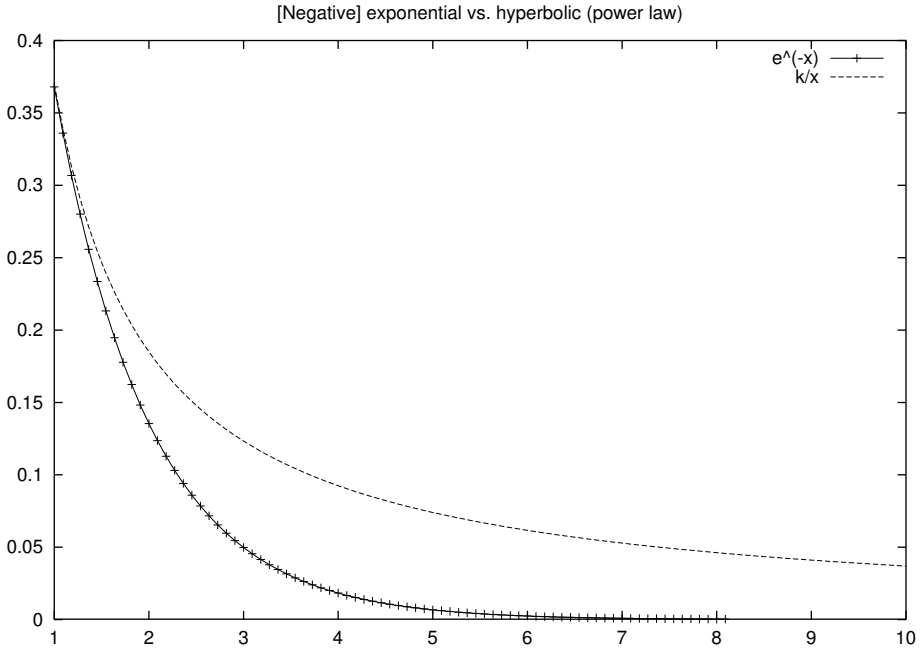
Word	Prob. (%)	Digram	Prob. (%)	Trigram	Prob. (%)
the	6.15	of the	0.95	one of the	0.03
of	3.54	in the	0.55	as well as	0.02
and	2.70	to the	0.33	the United States	0.02
to	2.51	on the	0.23	out of the	0.02
a	2.14	and the	0.21	some of the	0.02
in	1.90	for the	0.17	the end of	0.01
that	0.97	to be	0.16	the fact that	0.01
is	0.95	at the	0.15	part of the	0.01
was	0.94	with the	0.14	to be a	0.01
for	0.86	of a	0.14	of the United	0.01
with	0.68	that the	0.13	a number of	0.01
as	0.65	from the	0.13	end of the	0.01
he	0.65	by the	0.13	members of the	0.01
The	0.64	in a	0.13	in order to	0.01
his	0.63	as a	0.09	the use of	0.01
be	0.61	with a	0.09	that he had	0.01
on	0.61	is a	0.08	the number of	0.01
it	0.54	it is	0.08	most of the	0.01
had	0.50	of his	0.08	side of the	0.01
by	0.49	was a	0.08	that he was	0.01
at	0.49	is the	0.08	in front of	0.01
I	0.44	had been	0.07	and in the	0.01
not	0.41	for a	0.07	there is a	0.01
are	0.41	it was	0.07	of the most	0.01
from	0.41	he was	0.07	It was a	0.01
or	0.40	into the	0.07	One of the	0.01
have	0.38	as the	0.07	there was a	0.01
...	...	...	...	...	...
Number of units	100237		539929		884371
Entropy (bits/word)	11.47		6.06		2.01
Entropy (bits/letter)	1.94		1.03		0.34

(OVER)

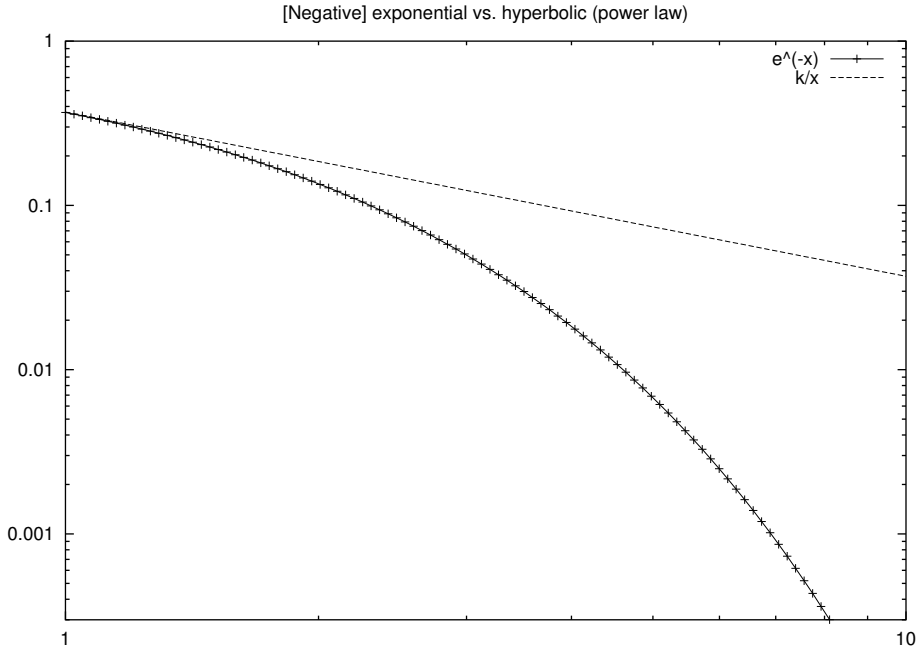
2. BCW. pg. 83


2000
1000
0
frequency
0
1000
2000
rank

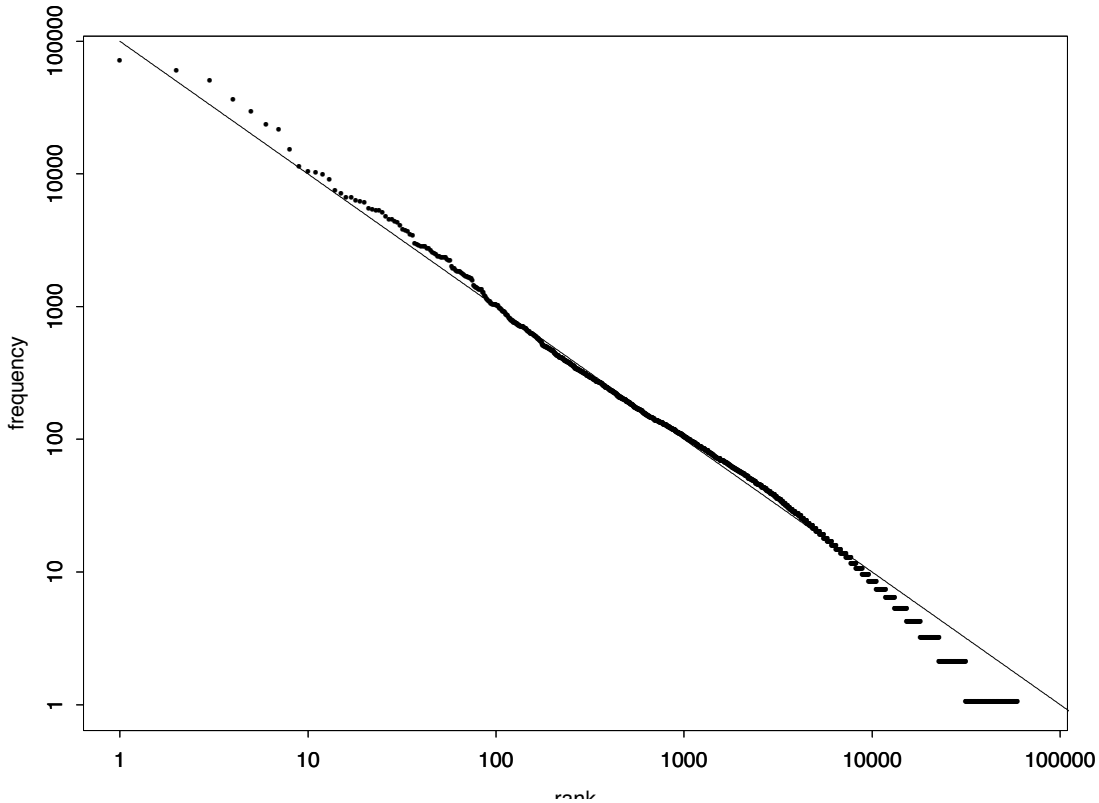
3.


[Negative] exponential vs. hyperbolic (power law)
e^(-x)
k/x
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9
10

4. Previous graph on a log-log scale



5. Manning and Schütze, pg 26: the Brown corpus, where the straight line is frequency $\times$ rank = 100,000.



6. BCW, pg. 80

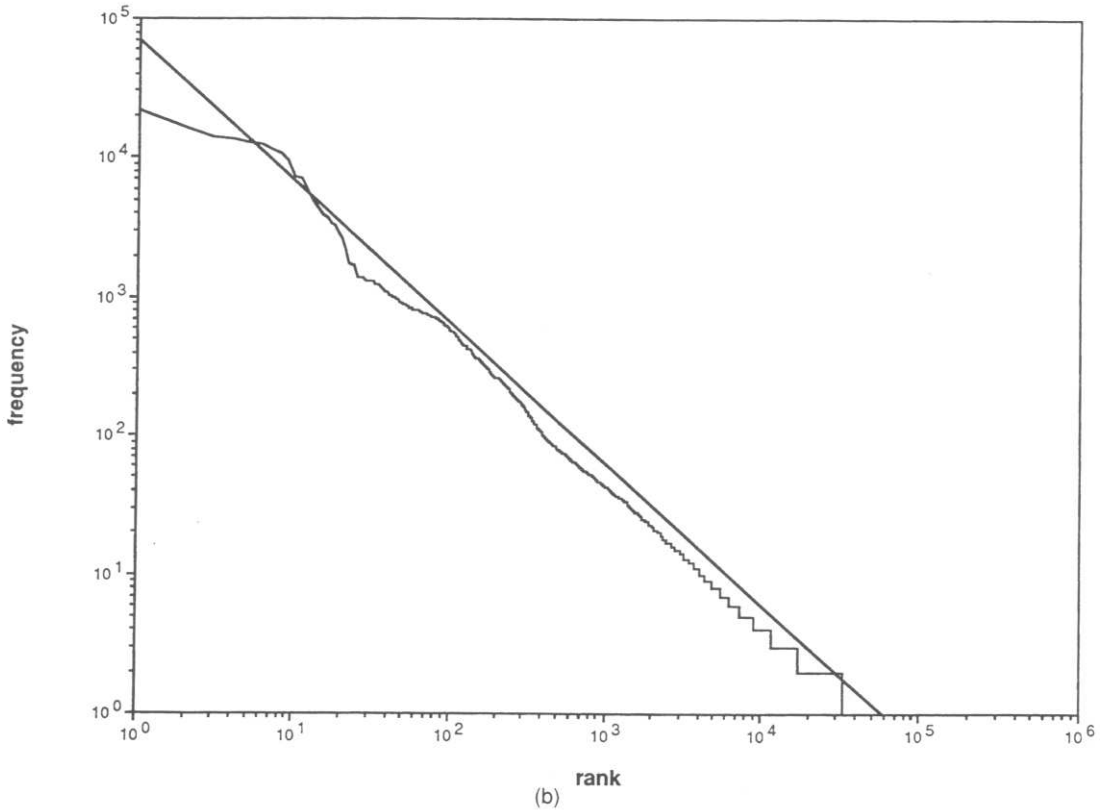
One striking way of illustrating the information content of such models is to generate text randomly according to them. Here are some characters chosen at random, where each has an equal chance of occurring:

)‘unHijz’YNvzweQSX,kjJRtylO’$(/~8}a"#\Dv*;-";^o.&uxPI)J‘XRfvt0uHIXegO)xZE&
vze"*&w#V[,;<(#v7Nm_1‘_x/ir$Ix6Ex8O~0lplyGDyOa+!/3zAs[U?EH]([sMo,{nXiy_
}A>2*~>F.RBi‘l?9\ !wd]&2M3IV&Mk eG>2R<Q2e>Ti8k)SHEeH<kt$9>[@&aZk(29
ti(OC\9uc]cF"ImZ5b^O;T*B5dH?wa3{!;L^3 U1w8W4bFnw(NGD"k 8QcWc__a\F@*‘
t;XIr(+8v>\E~:bk;zW9lUx,Oth05rpE.d(<lNU}kL^&gA,>VcW]Sj$"‘m20z? oE>xaEGQ
CN};Tevz#gxtEL_JNZR{jgU[,rn(75Zt}rLIXCgu+‘jj,JOu;,*$ae0nn9A.P>!{+sZ

This model has not captured any information about English text, and this is reflected by the gibberish produced! Even letters generated according to the order-0 statistics of Table 4-1 look more like English:

fsn’iaad ir lntns hynci,.aais oayimh t n ,at oeotc fheotyi t afrtgt oidtsO, wrr thraeoe
rdaFr ce.g psNo is.emahntawe,ei t etaodgdna- &em r n nd fih an f tpteaaInmas ss n
t"bar o be um oon tsrcs et mi ithyoitt h u ans w vsgr tn heaacrY.d erfdut y c, a,m
<hra Pieodn nyeSrsoto oea nlorseo j r s t w ge g E ikdeAJ .l eeTJiahednn ,ngaosl
dshoHo eh seelm G os threen nrgifeo,edsoht tgt n tiI a issnin"abi"h nht.e bs co
efhetntoilgevtnnadrtssaa ka dfnssiivb kuniseeaoM4l h acdchnr onoal ie a lhehtr
webYolo aere mblefeuom eomtlkIo h oattogodrinl aw Blbe.

7. BCW, pg 90



(b)

Figure 4-3(cont) (b) Rank-frequency graph for words in order-0 random text.