Common words in *Tom Sawyer*

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Frequencies of frequencies in *Tom Sawyer*

Word	Frequency of
Frequency	Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11–50	540
51–100	99
> 100	102

Zipf’s law in *Tom Sawyer*

Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
Oh	116	90	10440
two	104	100	10400

Word	Freq.	Rank	$f \cdot r$
	(f)	(r)	
turned	51	200	10200
you’ll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
Could	2	4000	8000
Applausive	1	8000	8000

Zipf’s law

$$f \propto \frac{1}{r} \tag{1}$$

There is a constant k such that

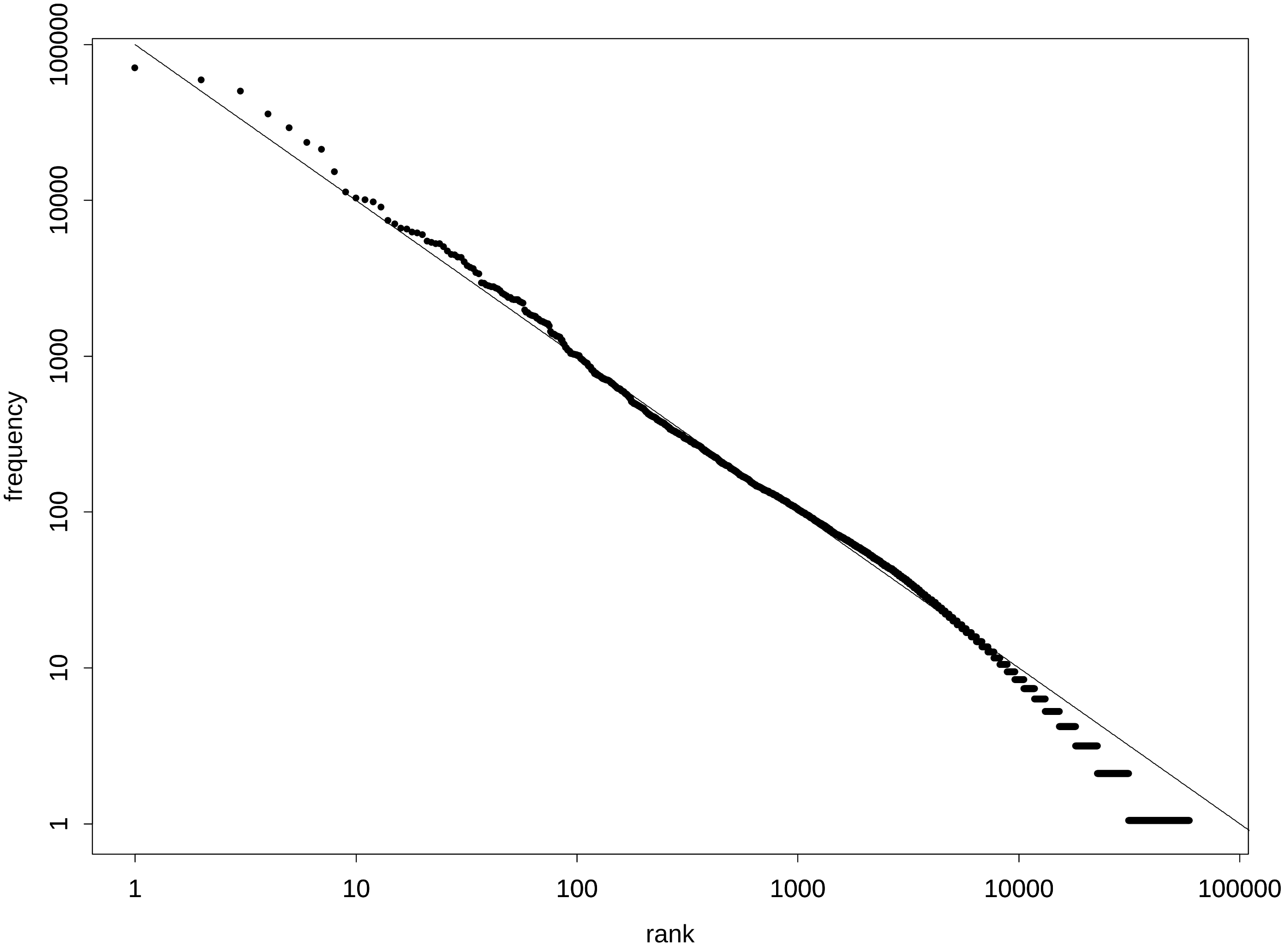
$$f \cdot r = k \tag{2}$$

Mandelbrot’s law

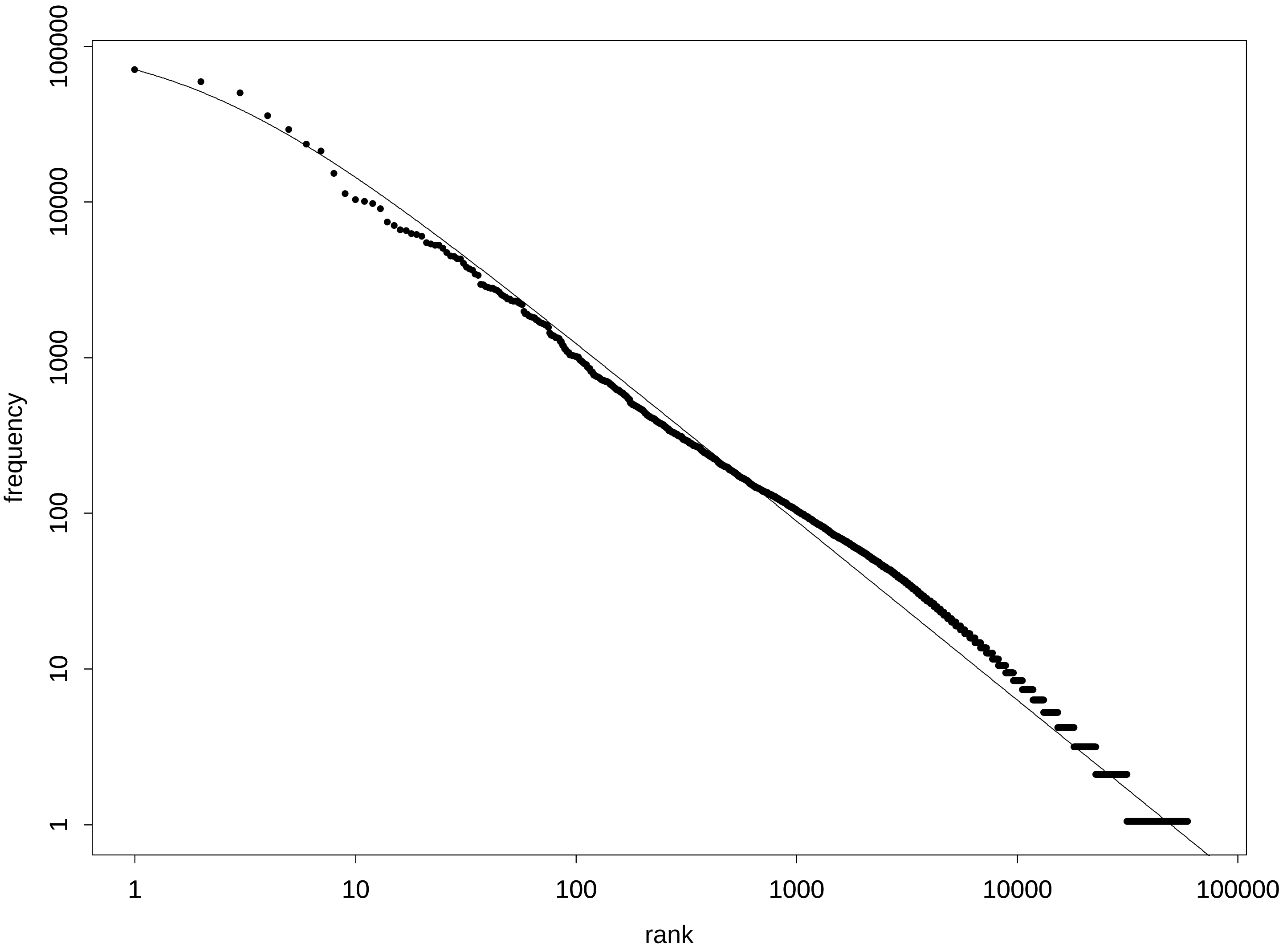
$$f = P(r + \rho)^{-B} \tag{3}$$

$$\log f = \log P - B \log(r + \rho) \tag{4}$$

Zipf’s law for the Brown corpus



Mandelbrot's formula for the Brown corpus



$P = 10^{5.4}$, $B = 1.15$, $\rho = 100$

Commonest bigrams in the *NYT*

Frequency	Word 1	Word 2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Filtered common bigrams in the *NYT*

Frequency	Word 1	Word 2	POS pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

KWIC display

```
 1   could find a target. The librarian  “showed  off” - running hither and thither w
 2  elights in. The young lady teachers  “showed  off” - bending sweetly over pupils
 3  ingly. The young gentlemen teachers  “showed  off” with small scoldings and other
 4  seeming vexation). The little girls  “showed  off” in various ways, and the littl
 5  n various ways, and the little boys  “showed  off” with such diligence that the a
 6  t genuwyne?” Tom lifted his lip and   showed  the vacancy. “Well, all right,” sai
 7  is little finger for a pen. Then he   showed  Huckleberry how to make an H and an
 8  ow’s face was haggard, and his eyes   showed  the fear that was upon him. When he
 9  not overlook the fact that Tom even   showed  a marked aversion to these inquests
10  own. Two or three glimmering lights   showed  where it lay, peacefully sleeping,
11  ird flash turned night into day and   showed  every little grass-blade, separate
12   that grew about their feet. And it   showed  three white, startled faces, too. A
13  he first thing his aunt said to him   showed  him that he had brought his sorrows
14  p from her lethargy of distress and   showed  good interest in the proceedings. S
15  ent a new burst of grief from Becky   showed  Tom that the thing in his mind had
16   shudder quiver all through him. He   showed  Huck the fragment of candle-wick pe
```

Syntactic frames for *showed* in *Tom Sawyer*

NP$_{agent}$ showed off (PP[*with/in*]$_{manner}$)

$$\text{NP}_{agent}\ \text{showed}\ (\text{NP}_{recipient})\ \left(\left\{\begin{array}{l} \text{NP}_{content} \\ \text{CP}[\textit{that}]_{content} \\ \text{VP}[\text{inf}]_{content} \\ \textit{how}\ \text{VP}[\text{inf}]_{content} \\ \text{CP}[\textit{where}]_{content} \end{array}\right\}\right)$$

NP$_{agent}$ showed NP[*interest*] PP[*in*]$_{content}$
NP$_{agent}$ showed NP[*aversion*] PP[*to*]$_{content}$