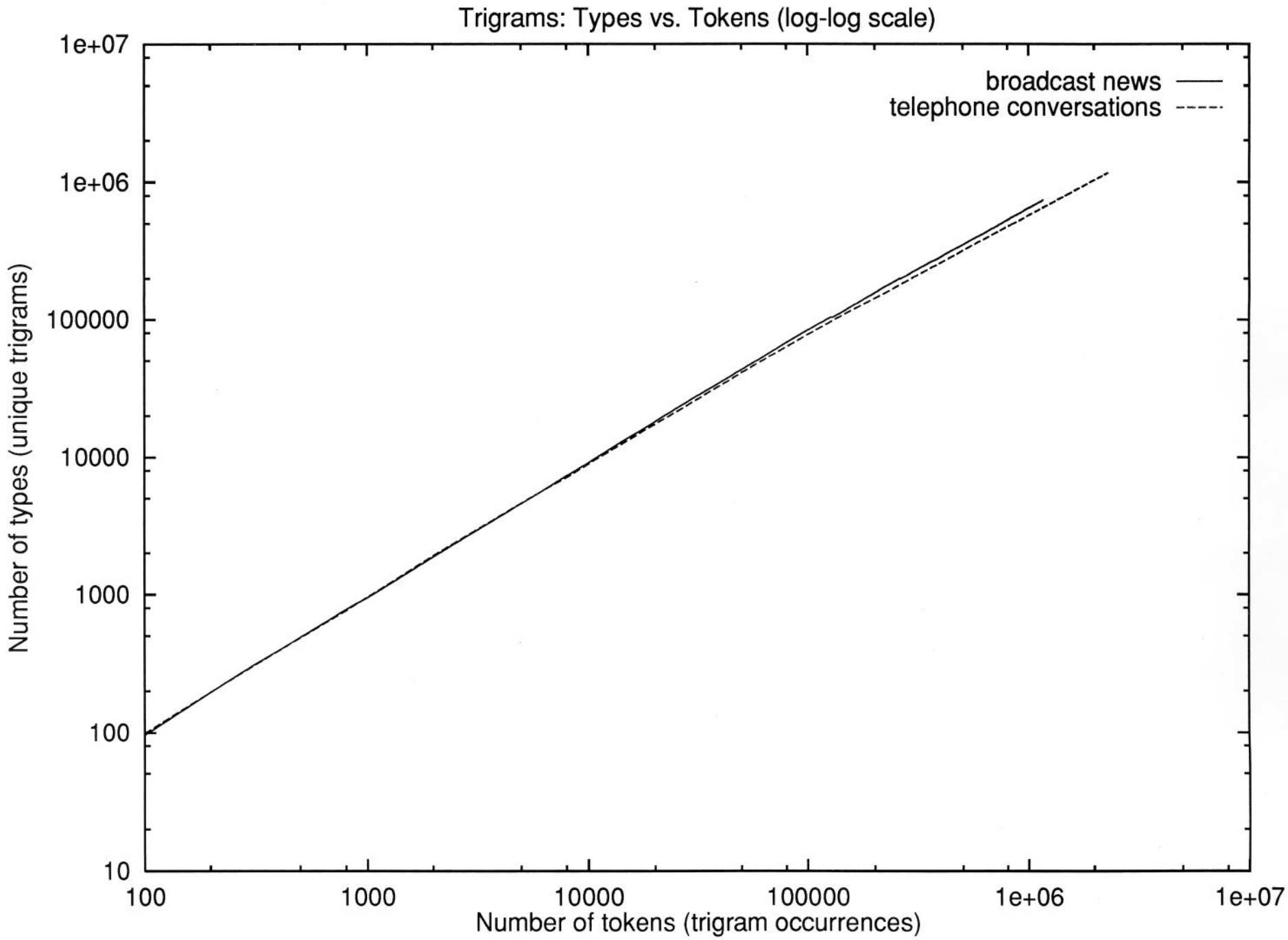
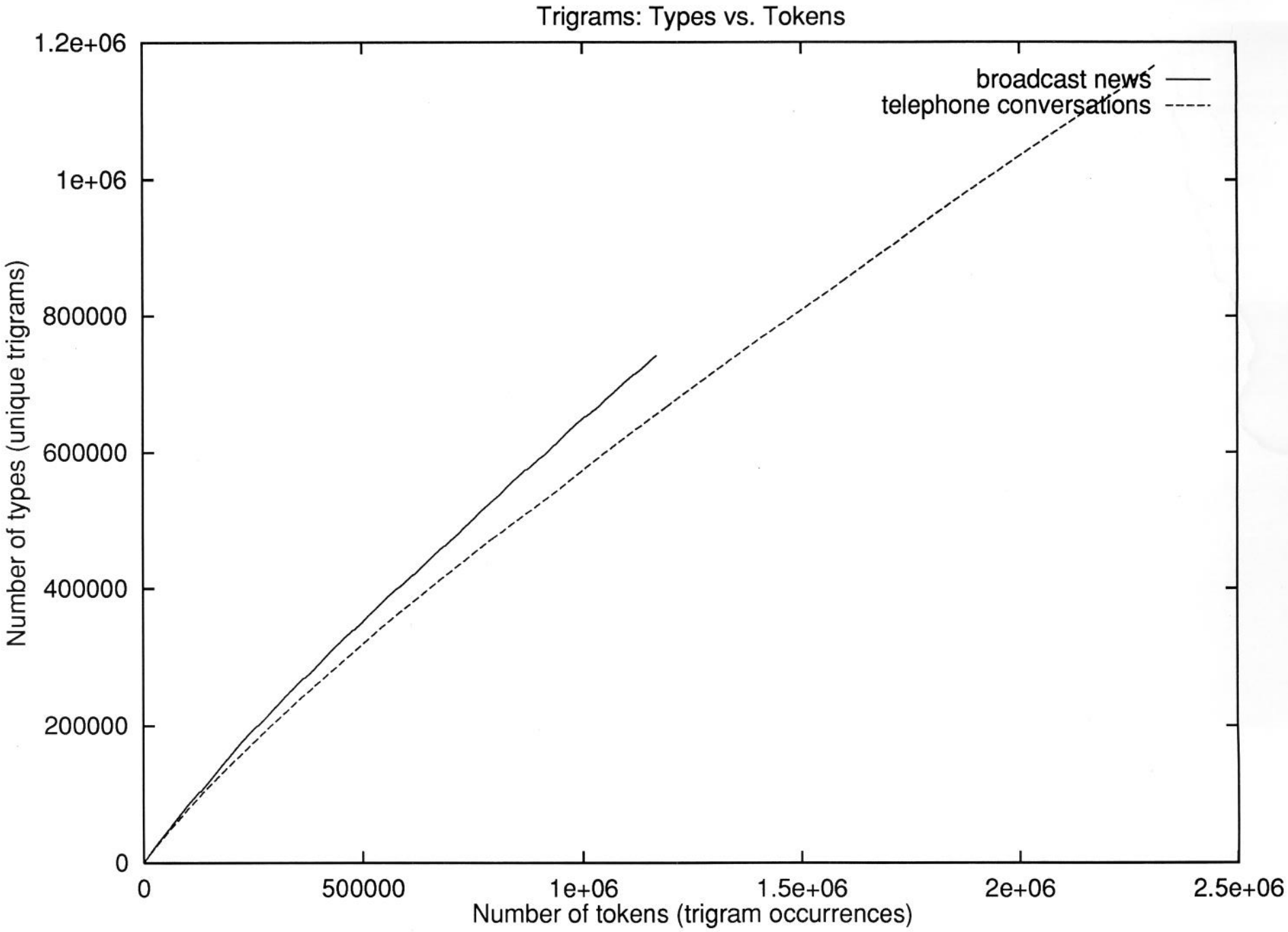
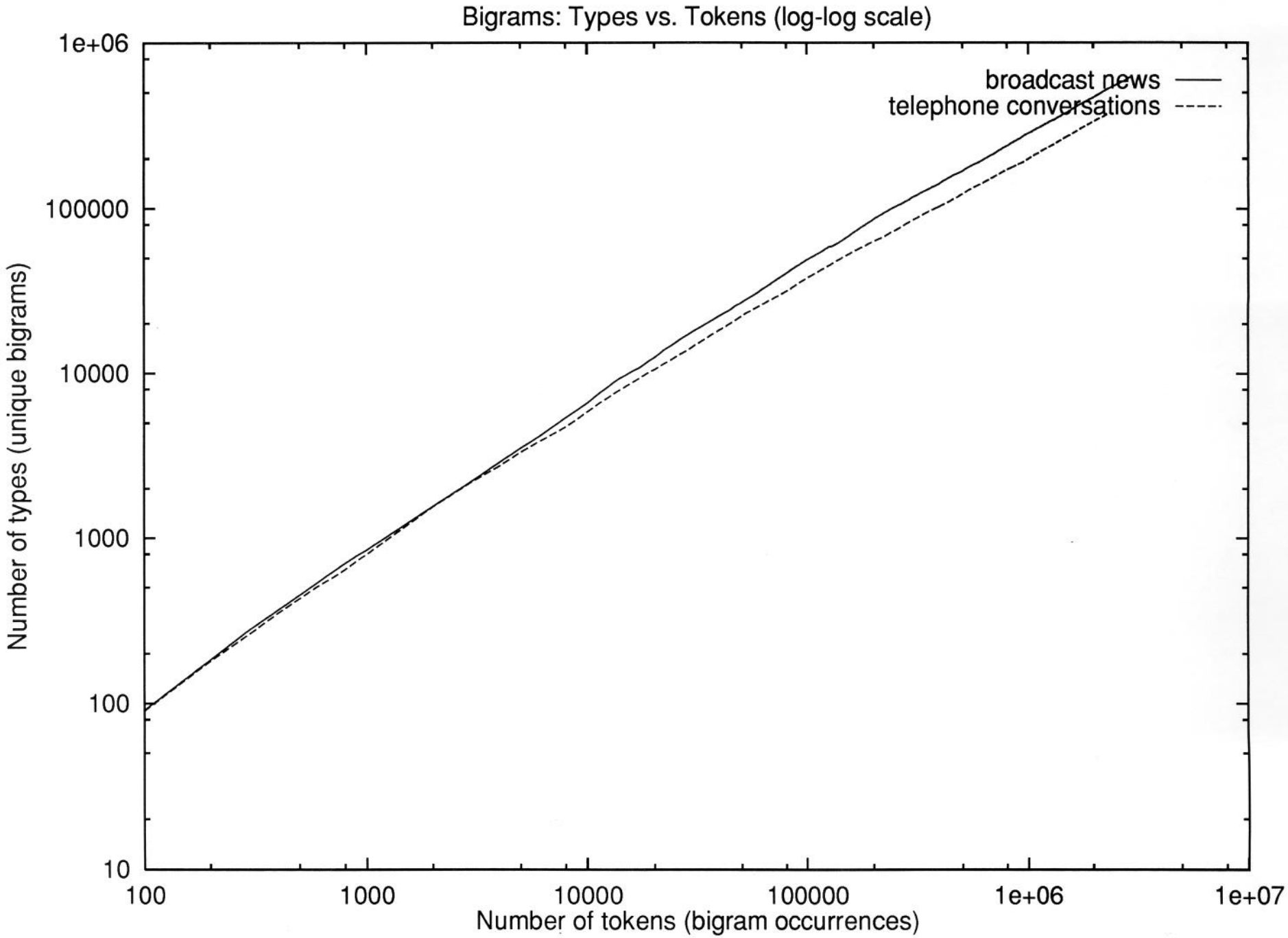
```
##    1000000 = total count for this file
THE     49214
</s>    42253
TO      24366
OF      24252
A       22499
AND     19474
IN      18954
THAT     9131
FOR      8994
IS       7707
ONE      7375
ON       5652
point    5483
AS       5473
SAID     5438
WITH     5365
IT       5332
FIVE     5168
TWO      5166
DOLLARS  5068
AT       4931
MR.      4850
BY       4813
FROM     4606
percent  4383
HUNDRED  4303
THREE    4119
HE       4032
ARE      3983
ITS      3867
BE       3806
WAS      3718
AN       3662
MILLION  3540
HAS      3522
HAVE     3503
BUT      3481
WILL     3405
...
...
ZELLERS     1
ZEMECKIS    1
ZEN         1
ZENITH'S    1
ZENKER      1
ZEOLITE     1
ZEOS'       1
ZEROED      1
ZEROS       1
ZESTY       1
ZEUS'S      1
ZHI         1
ZHONGTIAN   1
ZIGZAG      1
ZIGZAGGING  1
ZILLION     1
ZIONIST     1
ZIP         1
ZIPPER      1
ZIPPY       1
ZOO         1
ZOOKEEPER   1
```

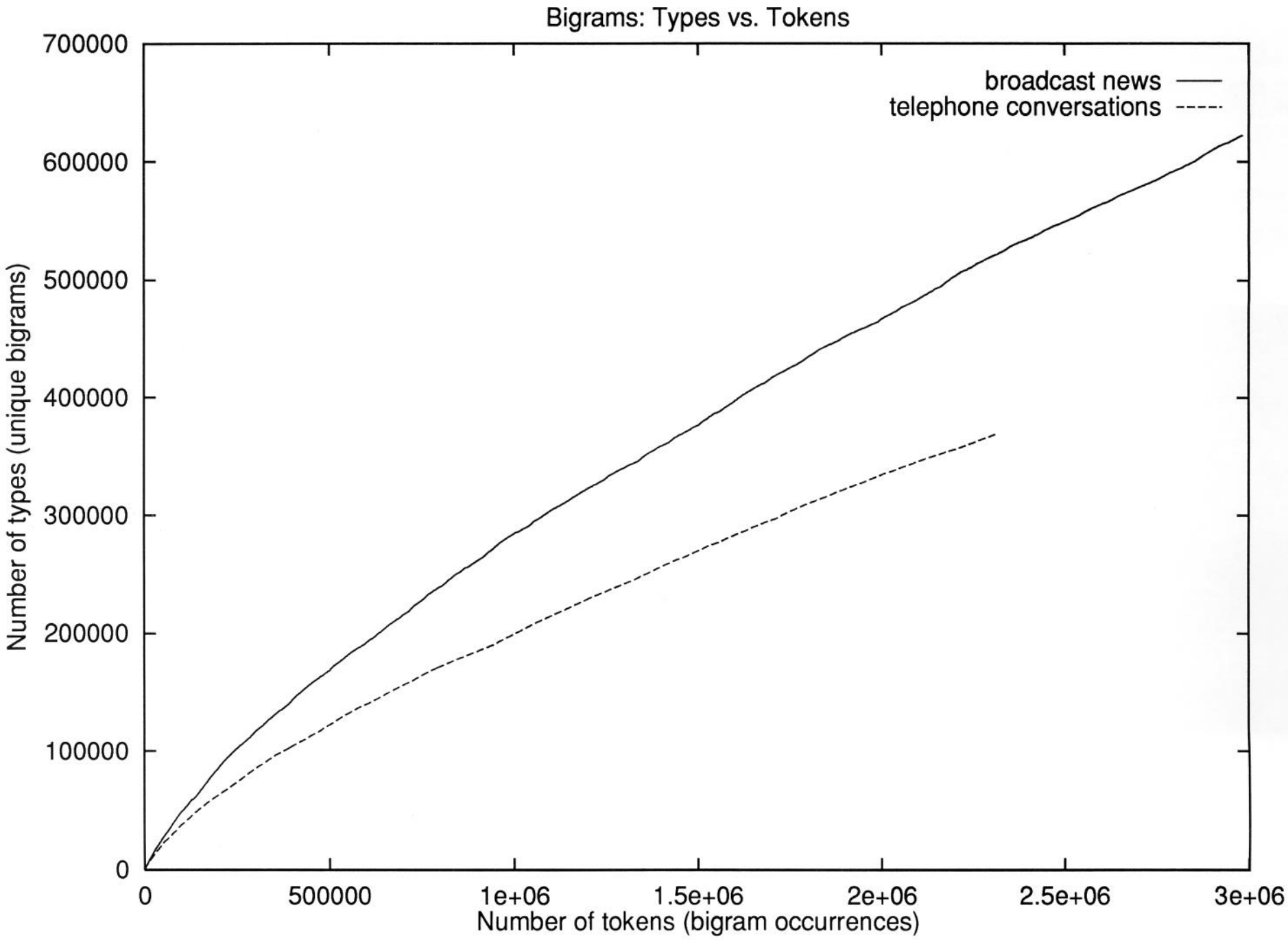
```
##    1000013 = total count for this file
</s>    62464
THE     48964
TO      26771
AND     24786
A       21747
OF      21359
IN      17071
THAT    16220
IS      12847
YOU     12042
I       11818
IT      10141
FOR      8145
THIS     7896
ON       6746
HAVE     6369
ARE      6182
WE       6061
THEY     5885
BE       5632
WITH     5502
BUT      5472
WAS      5162
HE       4722
AS       4666
NOT      4586
WHAT     4139
AT       4099
IT'S     4083
ONE      4013
ABOUT    3955
THINK    3724
SO       3699
THERE    3669
ALL      3530
DO       3524
FROM     3474
HAS      3208
...
...
ZEROS   1
ZHA     1
ZHIVAGOS        1
ZIANGSHING      1
ZILLIONS        1
ZIMBABWE'S      1
ZINGA   1
ZION    1
ZIONIST 1
ZOG     1
ZOIST   1
ZOO'S   1
ZOOMED  1
ZUCKERMAN       1
ZULU    1
ZUNG    1
ZURICH  1
ZWEIMAR 1
ZWICK'S 1
ZWINKELS        1
ZYGOTES 1
ZYUGANOV        1
```

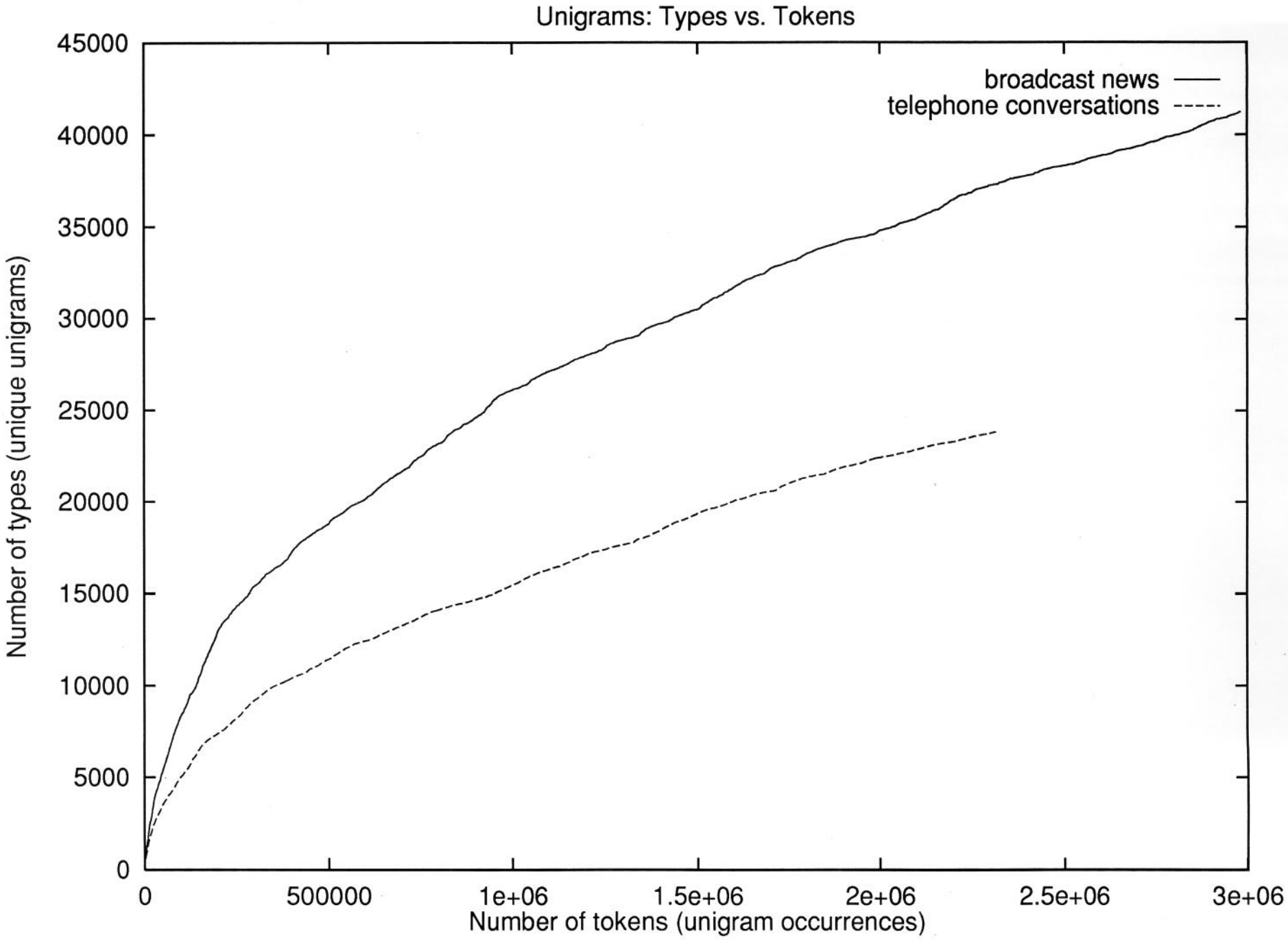
```
##    1000367 = total count for this file
I       38494
AND     33859
<SIL>   31220
THE     28067
YOU     25959
UH      25503
A       24431
TO      22812
THAT    20614
IT      17237
OF      16807
KNOW    15679
YEAH    14263
IN      12271
+NOISE+ 11792
THEY    10594
UH-HUH  10236
HAVE     9714
BUT      9011
SO       8977
IT'S     8108
IS       8016
WE       7923
WAS      7503
JUST     7027
WELL     6911
LIKE     6880
DO       6815
FOR      6477
THAT'S   6371
THINK    5876
DON'T    5809
OH       5647
RIGHT    5551
OR       5305
ON       5286
MY       5248
REALLY   4656
...
...
YEARBOOK        1
YEARS"  1
YELLER  1
YELLOWISH       1
YELLS   1
YIELD   1
YIP     1
YOGURT  1
YORKER  1
YOUNT   1
YOURSELFER      1
YUPPISH 1
ZACK    1
ZAK'S   1
ZALES   1
ZANTH   1
ZEALAND 1
ZEALOTS 1
ZEROED  1
ZIRCONIUHS      1
ZONES   1
__      1
```

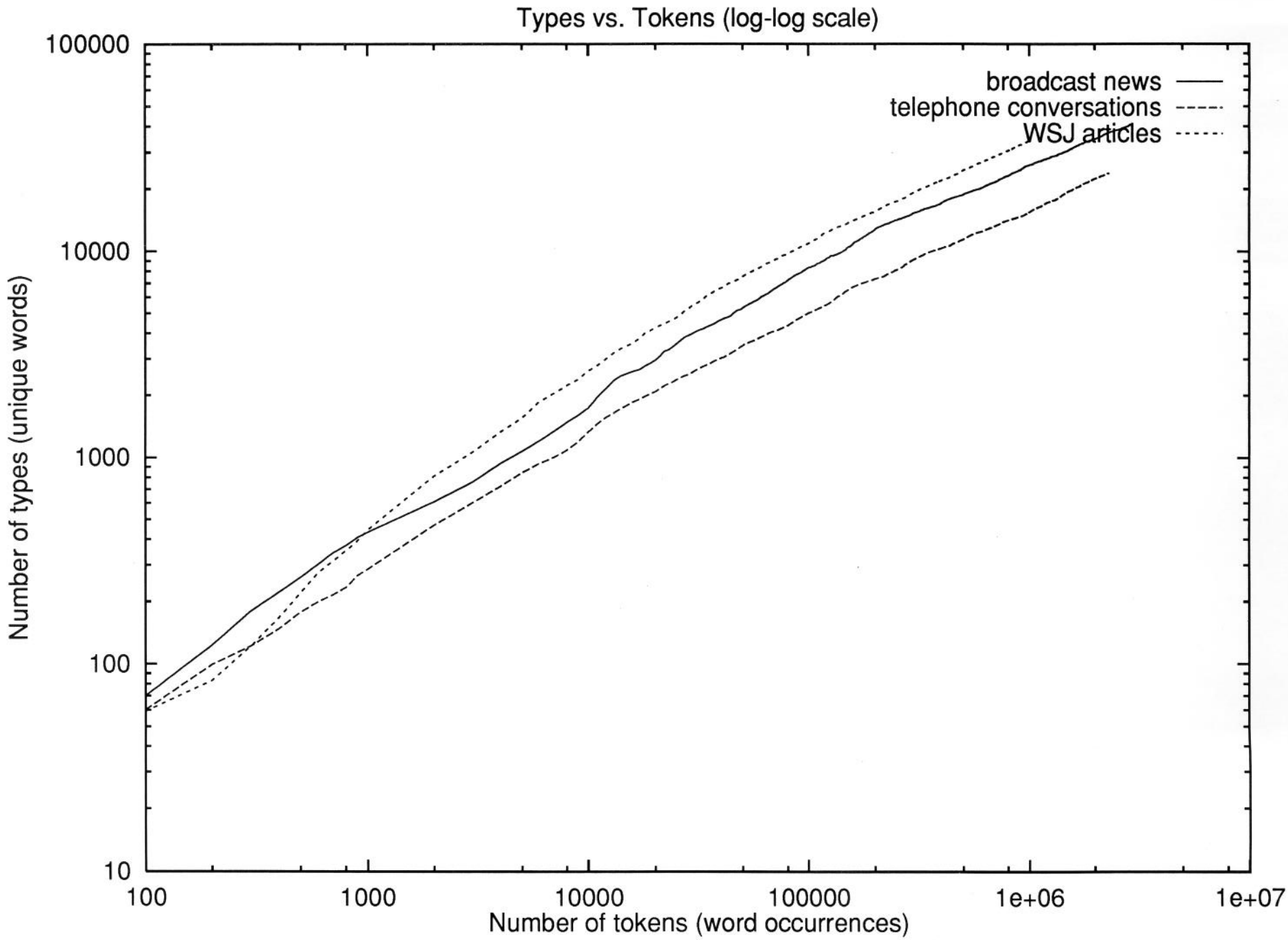

Trigrams: Types vs. Tokens (log-log scale)
broadcast news
telephone conversations
Number of types (unique trigrams)
Number of tokens (trigram occurrences)
1e+07
1e+06
100000
10000
1000
100
10
100
1000
10000
100000
1e+06
1e+07

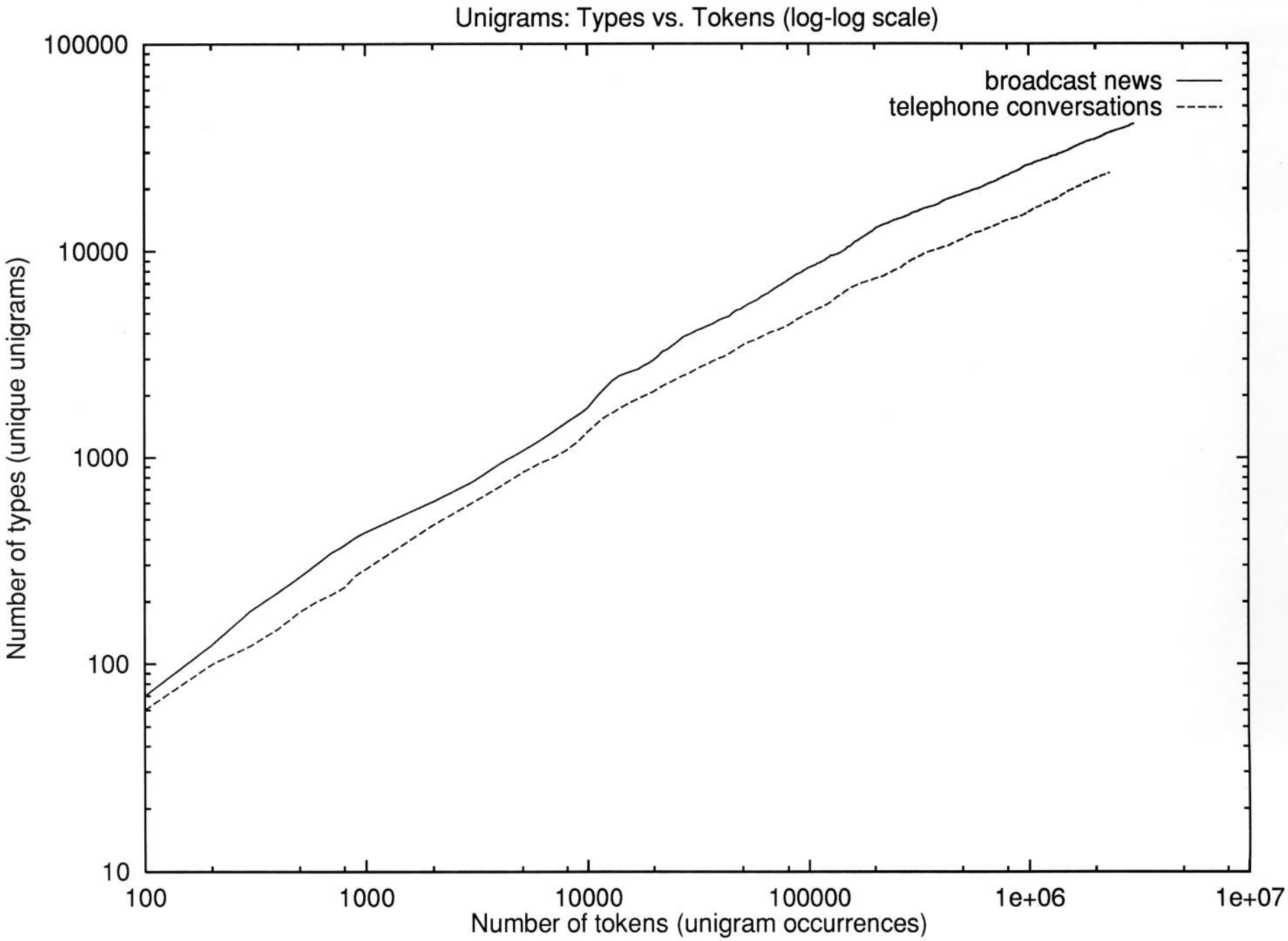

Trigrams: Types vs. Tokens
broadcast news
telephone conversations
Number of types (unique trigrams)
Number of tokens (trigram occurrences)
0
200000
400000
600000
800000
1e+06
1.2e+06
0
500000
1e+06
1.5e+06
2e+06
2.5e+06

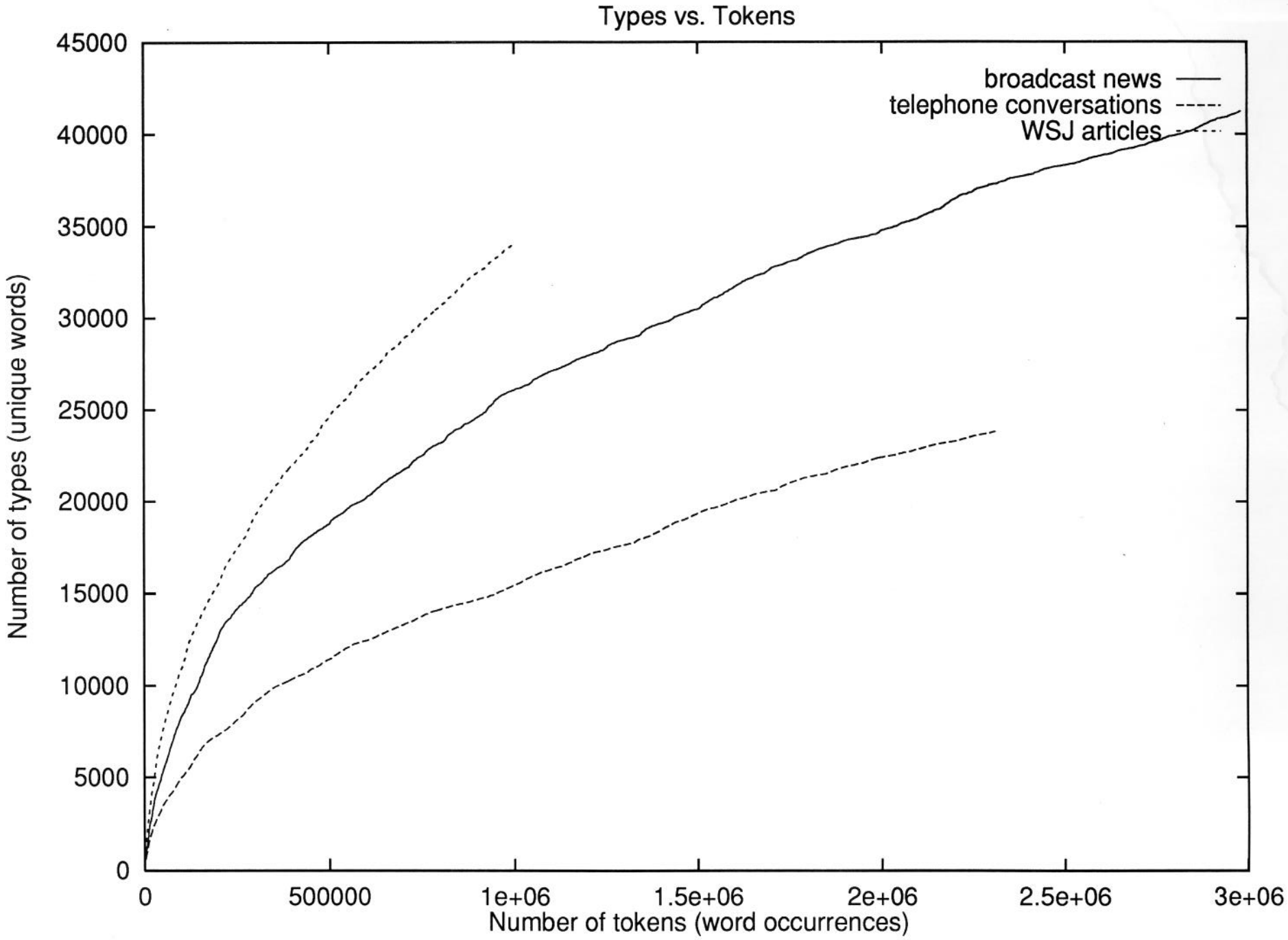

Bigrams: Types vs. Tokens (log-log scale)
broadcast news
telephone conversations
Number of types (unique bigrams)
Number of tokens (bigram occurrences)
1e+06
100000
10000
1000
100
10
100
1000
10000
100000
1e+06
1e+07


Bigrams: Types vs. Tokens
broadcast news
telephone conversations
Number of types (unique bigrams)
Number of tokens (bigram occurrences)
0
100000
200000
300000
400000
500000
600000
700000
0
500000
1e+06
1.5e+06
2e+06
2.5e+06
3e+06


Unigrams: Types vs. Tokens
broadcast news
telephone conversations
Number of types (unique unigrams)
Number of tokens (unigram occurrences)
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
0
500000
1e+06
1.5e+06
2e+06
2.5e+06
3e+06


Types vs. Tokens (log-log scale)
Number of types (unique words)
Number of tokens (word occurrences)
100000
10000
1000
100
10
100
1000
10000
100000
1e+06
1e+07
broadcast news
telephone conversations
WSJ articles


Unigrams: Types vs. Tokens (log-log scale)
broadcast news
telephone conversations
Number of types (unique unigrams)
Number of tokens (unigram occurrences)
100000
10000
1000
100
10
100
1000
10000
100000
1e+06
1e+07


Types vs. Tokens
broadcast news
telephone conversations
WSJ articles
Number of types (unique words)
Number of tokens (word occurrences)
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
0
500000
1e+06
1.5e+06
2e+06
2.5e+06
3e+06

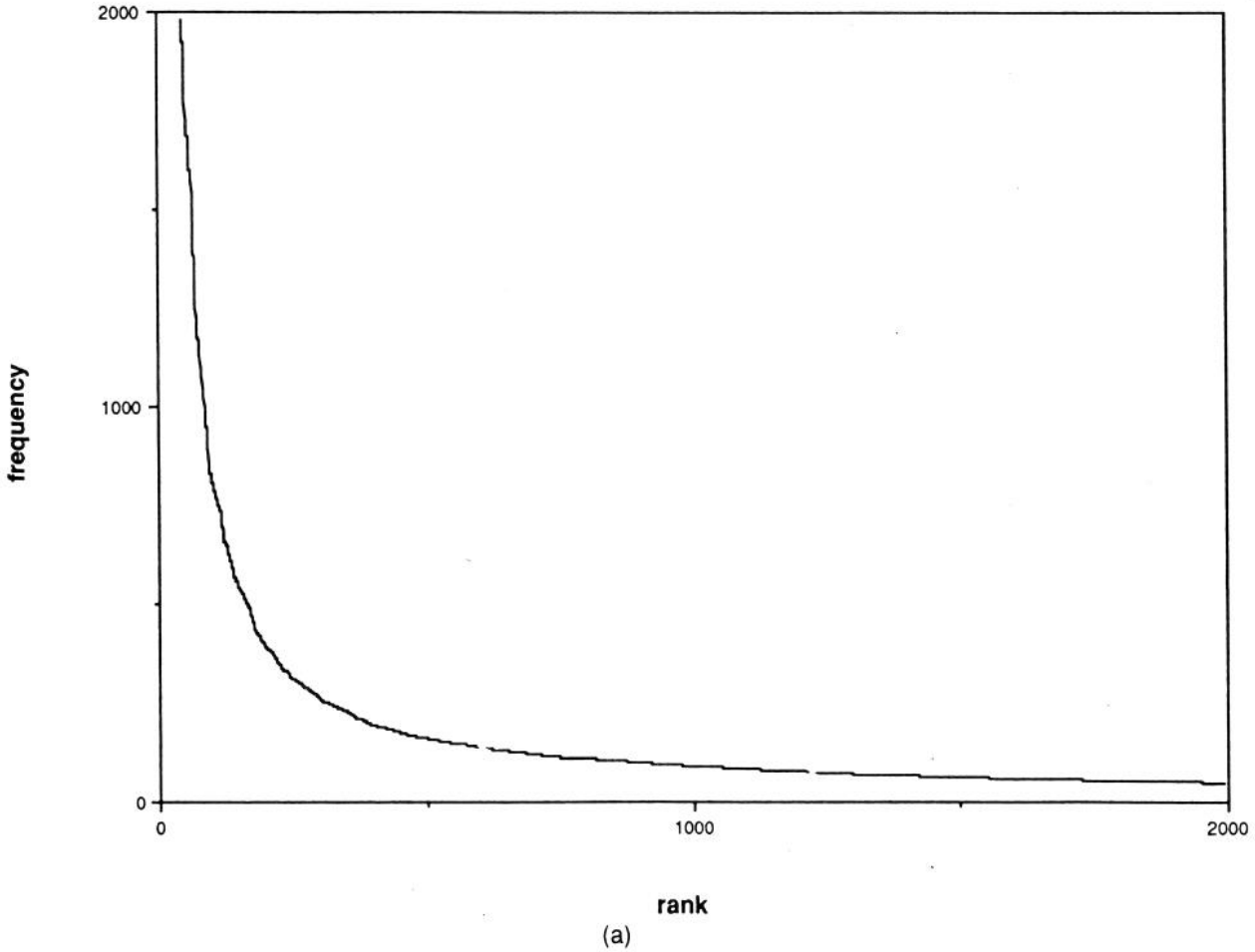
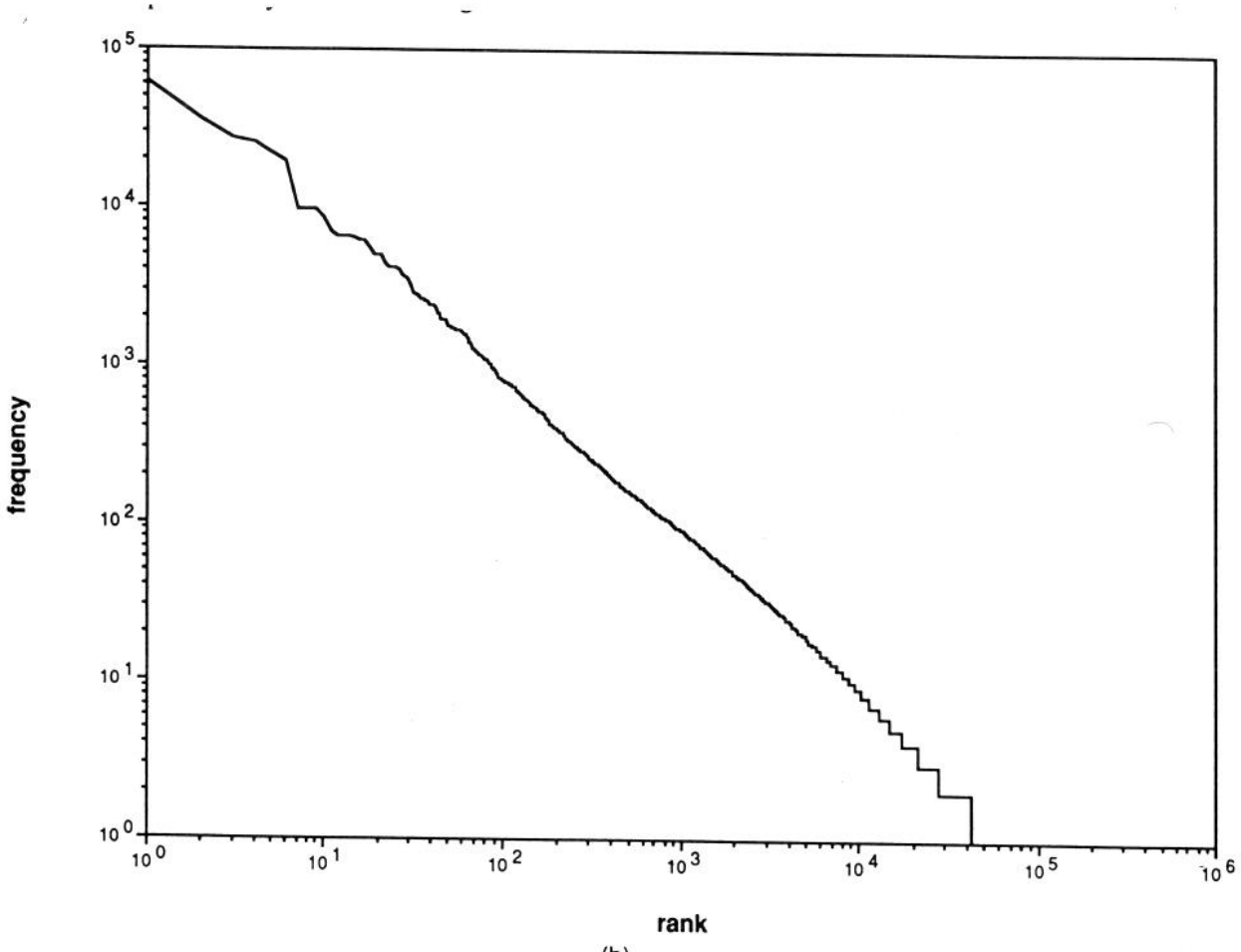


Figure 4-1 (a) Word-frequency data from the Brown corpus.



(b)

Figure 4-1(cont) (b) The same data plotted on logarithmic scales.

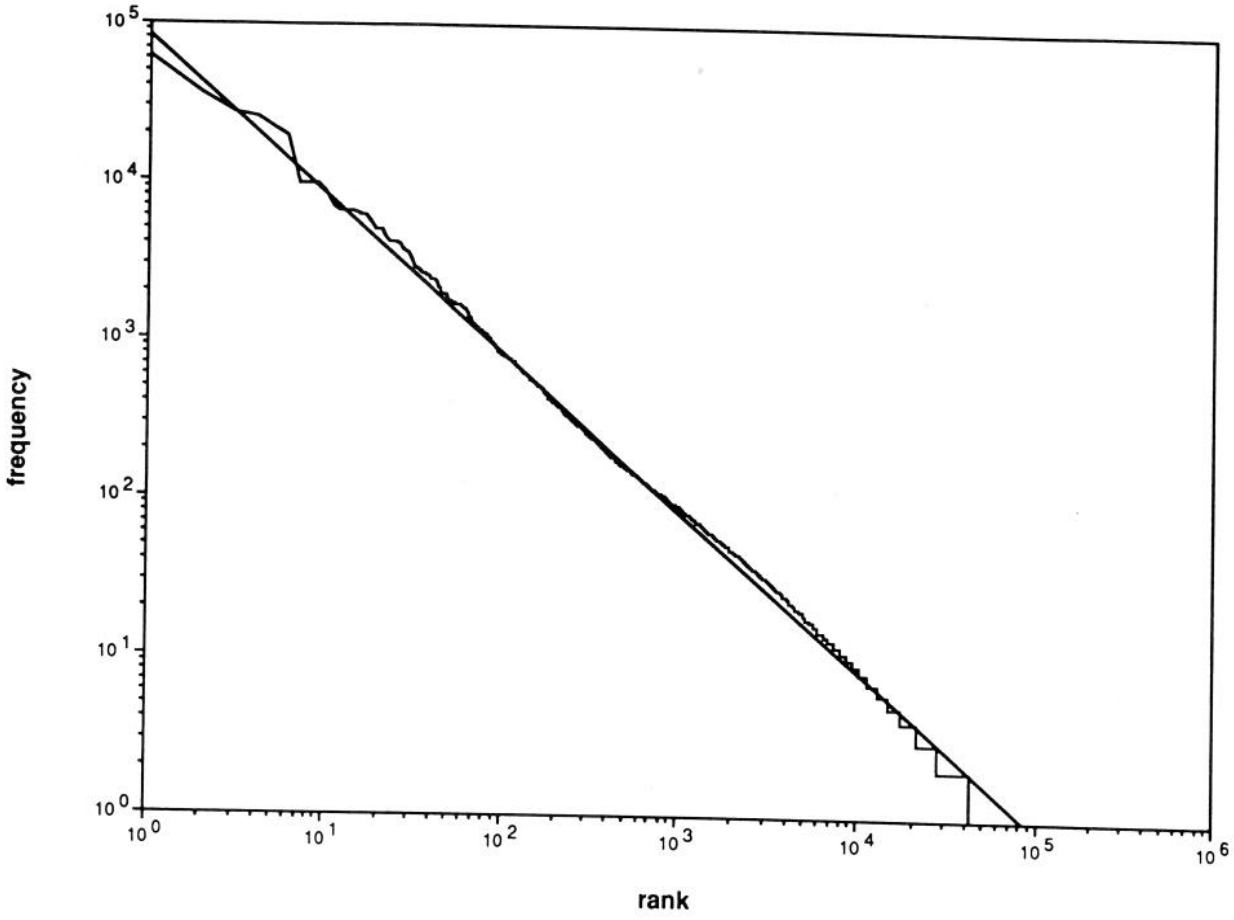


Figure 4-4 Word-frequency data from the Brown corpus with Zipf distribution (straight line).

TABLE 4-1 LETTER STATISTICS FROM THE BROWN CORPUS

	Letter	Prob. (%)	Digram	Prob. (%)	Trigram	Prob. (%)	Tetragram	Prob. (%)
	•	17.41	e•	3.05	•th	1.62	•the	1.25
	e	9.76	•t	2.40	the	1.36	the•	1.04
	t	7.01	th	2.03	he•	1.32	•of•	0.60
	a	6.15	he	1.97	•of	0.63	and•	0.48
	o	5.90	•a	1.75	of•	0.60	•and	0.46
	i	5.51	s•	1.75	ed•	0.60	•to•	0.42
	n	5.50	d•	1.56	•an	0.59	ing•	0.40
	s	4.97	in	1.44	nd•	0.57	•in•	0.32
	r	4.74	t•	1.38	and	0.55	tion	0.29
	h	4.15	n•	1.28	•in	0.51	n•th	0.23
	l	3.19	er	1.26	ing	0.50	f•th	0.21
	d	3.05	an	1.18	•to	0.50	of•t	0.21
	c	2.30	•o	1.14	to•	0.46	hat•	0.20
	u	2.10	re	1.10	ng•	0.44	•tha	0.20
	m	1.87	on	1.00	er•	0.39	.•••	0.20
	f	1.76	•s	0.99	in•	0.38	his•	0.19
	p	1.50	,•	0.96	is•	0.37	•for	0.19
	g	1.47	•i	0.93	ion	0.36	ion•	0.18
	w	1.38	•w	0.92	•a•	0.36	that	0.17
	y	1.33	at	0.87	on•	0.35	•was	0.17
	b	1.10	en	0.86	as•	0.33	d•th	0.16
	,	0.98	r•	0.83	•co	0.32	•is•	0.16
	.	0.83	y•	0.82	re•	0.32	was•	0.16
	v	0.77	nd	0.81	at•	0.31	t•th	0.16
	k	0.49	.•	0.81	ent	0.30	atio	0.15
	T	0.30	•h	0.78	e•t	0.30	•The	0.15
	"	0.29	ed	0.77	tio	0.29	e•th	0.15
	...	...	...	...	...	...	...	...
Number of units		94		3410		30249		131517
Entropy (bits/letter)		4.47		3.59		2.92		2.33

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