English Letter Frequency Counts:
Mayzner Revisited
or
ETAOIN SRHLDCU

Introduction

On December 17th 2012, I got a nice letter from Mark Mayzner, a retired 85-year-old researcher who studied the frequency of letter combinations in English words in the early 1960s. His 1965 publication has been cited in hundreds of articles. Mayzner describes his work:

> *I culled a corpus of 20,000 words from a variety of sources, e.g., newspapers, magazines, books, etc. For each source selected, a starting place was chosen at random. In proceeding forward from this point, all three, four, five, six, and seven-letter words were recorded until a total of 200 words had been selected. This procedure was duplicated 100 times, each time with a different source, thus yielding a grand total of 20,000 words. This sample broke down as follows: three-letter words, 6,807 tokens, 187 types; four-letter words, 5,456 tokens, 641 types; five-letter words, 3,422 tokens, 856 types; six-letter words, 2,264 tokens, 868 types; seven-letter words, 2,051 tokens, 924 types. I then proceeded to construct tables that showed the frequency counts for three, four, five, six, and seven-letter words, but most importantly, broken down by word length and letter position, which had never been done before to my knowledge.*

and he wonders if:

> *perhaps your group at Google might be interested in using the computing power that is now available to significantly expand and produce such tables as I constructed some 50 years ago, but now using the Google Corpus Data, not the tiny 20,000 word sample that I used.*

The answer is: yes indeed, I am interested! And it will be a lot easier for me than it was for Mayzner. Working 60s-style, Mayzner had to gather his collection of text sources, then go through them and select individual words, punch them on Hollerith cards, and use a card-sorting machine.

Here's what we can do with today's computing power (using publicly available data and the processing power of my own personal computer; I'm not relying on access to corporate computing power):

1. I consulted the Google books Ngrams raw data set, which gives word counts of the number of times each word is mentioned (broken down by year of publication) in the books that have been scanned by Google.
2. I downloaded the English Version 20120701 "1-grams" (that is, word counts) from that data set given as the files "a" to "z" (that is, http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-a.gz to http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-1gram-20120701-z.gz). I unzipped each file; the result is 23 GB of text (so don't try to download them on your

phone).

3. I then condensed these entries, combining the counts for all years, and for different capitalizations: "word", "Word" and "WORD" were all recorded under "WORD". I discarded any entry that used a character other than the 26 letters A-Z. I also discarded any word with fewer than 100,000 mentions. (If you want you can download the word count file; note that it is 1.5 MB.)
4. I generated tables of counts, first for words, then for letters and letter sequences, keyed off of the positions and word lengths.

Word Counts

My distillation of the Google books data gives us 97,565 distinct words, which were mentioned 743,842,922,321 times (37 million times more than in Mayzner's 20,000-mention collection). Each distinct word is called a "type" and each mention is called a "token." To no surprise, the most common word is "the". Here are the top 50 words, with their counts (in billions of mentions) and their overall percentage (looking like a Zipf distribution):

WORD	COUNT	PERCENT	bar graph
the	53.10 B	7.14%	the
of	30.97 B	4.16%	of
and	22.63 B	3.04%	and
to	19.35 B	2.60%	to
in	16.89 B	2.27%	in
a	15.31 B	2.06%	a
is	8.38 B	1.13%	is
that	8.00 B	1.08%	that
for	6.55 B	0.88%	for
it	5.74 B	0.77%	it
as	5.70 B	0.77%	as
was	5.50 B	0.74%	was
with	5.18 B	0.70%	with
be	4.82 B	0.65%	be
by	4.70 B	0.63%	by
on	4.59 B	0.62%	on
not	4.52 B	0.61%	not
he	4.11 B	0.55%	he
i	3.88 B	0.52%	i
this	3.83 B	0.51%	this
are	3.70 B	0.50%	are
or	3.67 B	0.49%	or
his	3.61 B	0.49%	his
from	3.47 B	0.47%	from
at	3.41 B	0.46%	at
which	3.14 B	0.42%	which
but	2.79 B	0.38%	but
have	2.78 B	0.37%	have
an	2.73 B	0.37%	an
had	2.62 B	0.35%	had
they	2.46 B	0.33%	they
you	2.34 B	0.31%	you

were	2.27 B	0.31%	were
their	2.15 B	0.29%	their
one	2.15 B	0.29%	one
all	2.06 B	0.28%	all
we	2.06 B	0.28%	we
can	1.67 B	0.22%	can
her	1.63 B	0.22%	her
has	1.63 B	0.22%	has
there	1.62 B	0.22%	there
been	1.62 B	0.22%	been
if	1.56 B	0.21%	if
more	1.55 B	0.21%	more
when	1.52 B	0.20%	when
will	1.49 B	0.20%	will
would	1.47 B	0.20%	would
who	1.46 B	0.20%	who
so	1.45 B	0.19%	so
no	1.40 B	0.19%	no

Word Lengths

And here is the breakdown of mentions (in millions) by word length (looking like a Poisson distribution). The average is 4.79 letters per word, and 80% are between 2 and 7 letters long:

LEN	COUNT	PERCENT	bar graph
1	22301.22 M	2.998%	1
2	131293.85 M	17.651%	2
3	152568.38 M	20.511%	3
4	109988.33 M	14.787%	4
5	79589.32 M	10.700%	5
6	62391.21 M	8.388%	6
7	59052.66 M	7.939%	7
8	44207.29 M	5.943%	8
9	33006.93 M	4.437%	9
10	22883.84 M	3.076%	10
11	13098.06 M	1.761%	11
12	7124.15 M	0.958%	12
13	3850.58 M	0.518%	13
14	1653.08 M	0.222%	14
15	565.24 M	0.076%	15
16	151.22 M	0.020%	16
17	72.81 M	0.010%	17
18	28.62 M	0.004%	18
19	8.51 M	0.001%	19
20	6.35 M	0.001%	20
21	0.13 M	0.000%	21
22	0.81 M	0.000%	22
23	0.32 M	0.000%	23

Here is the distribution for distinct words (that is, counting each word only once regardless of how many times it is mentioned). Now the average is 7.60 letters long, and 80% are between 4 and 10 letters long:

LEN	COUNT	PERCENT	bar graph
1	26	0.027%	1
2	662	0.679%	2
3	4,615	4.730%	3
4	6,977	7.151%	4
5	10,541	10.804%	5
6	13,341	13.674%	6
7	14,392	14.751%	7
8	13,284	13.616%	8
9	11,079	11.356%	9
10	8,468	8.679%	10
11	5,769	5.913%	11
12	3,700	3.792%	12
13	2,272	2.329%	13
14	1,202	1.232%	14
15	668	0.685%	15
16	283	0.290%	16
17	158	0.162%	17
18	64	0.066%	18
19	40	0.041%	19
20	16	0.016%	20
21	1	0.001%	21
22	5	0.005%	22
23	2	0.002%	23

Here are the 24 words with length of 20 or more (that are mentioned at least 100,000 times each in the book corpus):

```
electroencephalographic          radiopharmaceuticals
polytetrafluoroethylene          electroencephalogram
forschungsgemeinschaft           keratoconjunctivitis
deinstitutionalization           counterrevolutionary
counterrevolutionaries           immunohistochemistry
dehydroepiandrosterone           internationalisation
electroencephalography           hypercholesterolemia
immunoelectrophoresis            phosphatidylinositol
institutionalisation             compartmentalization
acetylcholinesterase             electrophysiological
internationalization             electrocardiographic
institutionalization             uncharacteristically
```

Letter Counts

Enough of words; let's get back to Mayzner's request and look at letter counts. There were 3,563,505,777,820 letters mentioned. Here they are in frequency order:

LET	COUNT	PERCENT	bar graph
E	445.2 B	12.49%	E
T	330.5 B	9.28%	T
A	286.5 B	8.04%	A
O	272.3 B	7.64%	O
I	269.7 B	7.57%	I

```
N 257.8 B   7.23%  N
S 232.1 B   6.51%  S
R 223.8 B   6.28%  R
H 180.1 B   5.05%  H
L 145.0 B   4.07%  L
D 136.0 B   3.82%  D
C 119.2 B   3.34%  C
U  97.3 B   2.73%  U
M  89.5 B   2.51%  M
F  85.6 B   2.40%  F
P  76.1 B   2.14%  P
G  66.6 B   1.87%  G
W  59.7 B   1.68%  W
Y  59.3 B   1.66%  Y
B  52.9 B   1.48%  B
V  37.5 B   1.05%  V
K  19.3 B   0.54%  K
X   8.4 B   0.23%  X
J   5.7 B   0.16%  J
Q   4.3 B   0.12%  Q
Z   3.2 B   0.09%  Z
```

Note there is a standard order of frequency used by typesetters, ETAOIN SHRDLU, that is slightly violated here: L, R, and C have all moved up one rank, giving us the less mnemonic ETAOIN SRHLDCU.

In the colored-bar chart below (inspired by the Wikipedia article on Letter Frequency), the frequency of each letter is proportional to the length of the color bar. If you hover the mouse over each color bar, you can see the exact percentages and counts. (This is the same information as in the table above, presented in a different way.)

e t a o i n s r h l d c
u m f p g w y b v k x j q z

Letter Counts by Position Within Word

Now we show the letter frequencies by position within word. That is, the frequencies for just the first letter in each word, just the second letter, and so on. We also show frequencies for positions relative to the end of the word: "-1" means the last letter, "-2" means the second to last, and so on. We can see that the frequencies vary quite a bit; for example, "e" is uncommon as the first letter (4 times less frequent than elsewhere); similarly "n" is 3 times less common as the first letter than it is overall. The letter "e" makes a comeback as the most common last letter (and also very common at 3rd and 5th letter places). The most common first letter is "t" and the most common second letter is "o".

1 e t a o i n s r h l d c u m
f p g w y b v k x j q z
2 e t a o i n s r h l
d c u m f p g w y b v k x j q z

3 e t a o i n s r h l d c u m f p g w y b v k x j q z

4 e t a o i n s r h l d c u m f p g w y b v k x j q z

5 e t a o i n s r h l d c u m f p g w y b v k x j q z

6 e t a o i n s r h l d c u m f p g w y b v k x j q z

7 e t a o i n s r h l d c u m f p g w y b v k x j q z

-7 e t a o i n s r h l d c u m f p g w y b v k x j q z

-6 e t a o i n s r h l d c u m f p g w y b v k x j q z

-5 e t a o i n s r h l d c u m f p g w y b v k x j q z

-4 e t a o i n s r h l d c u m f p g w y b v k x j q z

-3 e t a o i n s r h l d c u m f p g w y b v k x j q z

-2 e t a o i n s r h l d c u m f p g w y b v k x j q z

-1 e t a o i n s r h l d c u m f p g w y b v k x j q z

Two-Letter Sequence (Bigram) Counts

Now we turn to sequences of letters: consecutive letters anywhere within a word. In the list below are the 50 most frequent two-letter sequences (which are called "bigrams"):

BI	COUNT	PERCENT	bar graph
TH	100.3 B	(3.56%)	TH
HE	86.7 B	(3.07%)	HE
IN	68.6 B	(2.43%)	IN
ER	57.8 B	(2.05%)	ER
AN	56.0 B	(1.99%)	AN
RE	52.3 B	(1.85%)	RE
ON	49.6 B	(1.76%)	ON
AT	41.9 B	(1.49%)	AT
EN	41.0 B	(1.45%)	EN
ND	38.1 B	(1.35%)	ND
TI	37.9 B	(1.34%)	TI
ES	37.8 B	(1.34%)	ES

OR	36.0 B	(1.28%)	OR
TE	34.0 B	(1.20%)	TE
OF	33.1 B	(1.17%)	OF
ED	32.9 B	(1.17%)	ED
IS	31.8 B	(1.13%)	IS
IT	31.7 B	(1.12%)	IT
AL	30.7 B	(1.09%)	AL
AR	30.3 B	(1.07%)	AR
ST	29.7 B	(1.05%)	ST
TO	29.4 B	(1.04%)	TO
NT	29.4 B	(1.04%)	NT
NG	26.9 B	(0.95%)	NG
SE	26.3 B	(0.93%)	SE
HA	26.1 B	(0.93%)	HA
AS	24.6 B	(0.87%)	AS
OU	24.5 B	(0.87%)	OU
IO	23.5 B	(0.83%)	IO
LE	23.4 B	(0.83%)	LE
VE	23.3 B	(0.83%)	VE
CO	22.4 B	(0.79%)	CO
ME	22.4 B	(0.79%)	ME
DE	21.6 B	(0.76%)	DE
HI	21.5 B	(0.76%)	HI
RI	20.5 B	(0.73%)	RI
RO	20.5 B	(0.73%)	RO
IC	19.7 B	(0.70%)	IC
NE	19.5 B	(0.69%)	NE
EA	19.4 B	(0.69%)	EA
RA	19.3 B	(0.69%)	RA
CE	18.4 B	(0.65%)	CE
LI	17.6 B	(0.62%)	LI
CH	16.9 B	(0.60%)	CH
LL	16.3 B	(0.58%)	LL
BE	16.2 B	(0.58%)	BE
MA	15.9 B	(0.57%)	MA
SI	15.5 B	(0.55%)	SI
OM	15.4 B	(0.55%)	OM
UR	15.3 B	(0.54%)	UR

Below is a table of all 26 × 26 = 676 bigrams; in each cell the orange bar is proportional to the frequency, and if you hover you can see the exact counts and percentage. There are only seven bigrams that do not occur among the 2.8 trillion mentions: JQ, QG, QK, QY, QZ, WQ, and WZ. If you look closely you see they are shown as deleted.

AA	BA	CA	DA	EA
AB	BB	CB	DB	EB
AC	BC	CC	DC	EC
AD	BD	CD	DD	ED
AE	BE	CE	DE	EE
AF	BF	CF	DF	EF

AG	BG	CG	DG	EG
AH	BH	CH	DH	EH
AI	BI	CI	DI	EI
AJ	BJ	CJ	DJ	EJ
AK	BK	CK	DK	EK
AL	BL	CL	DL	EL
AM	BM	CM	DM	EM
AN	BN	CN	DN	EN
AO	BO	CO	DO	EO
AP	BP	CP	DP	EP
AQ	BQ	CQ	DQ	EQ
AR	BR	CR	DR	ER
AS	BS	CS	DS	ES
AT	BT	CT	DT	ET
AU	BU	CU	DU	EU
AV	BV	CV	DV	EV
AW	BW	CW	DW	EW
AX	BX	CX	DX	EX
AY	BY	CY	DY	EY
AZ	BZ	CZ	DZ	EZ

N-Letter Sequences (N-grams)

What are the most common n-letter sequences (called "n-grams") for various values of n? You can see the 50 most common for each value of n from 1 to 9 in the table below. The counts and percentages are not shown, but don't worry -- you'll get lots of counts in the next section.

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
	as	thi	were	rough	action	ulation	mportant	niversity
	ou	wit	tive	ative	person	another	rticular	therefore
	io	are	over	prese	eneral	importa	particul	nstitutio
	le	ess	ding	feren	system	interes	epresent	ification
	ve	not	pres	hough	relati	nterest	represen	establish
	co	ive	nter	ution	ctions	elation	increase	understan
	me	was	comp	roduc	ecause	rmation	individu	nderstand
	de	ect	able	resen	becaus	mportan	ndividua	difficult
	hi	rea	heir	thoug	before	product	dividual	structure
	ri	com	thei	press	ession	formati	elations	knowledge
	ro	eve	ally	first	develo	communi	nformati	struction
	ic	per	ated	after	evelop	lations	politica	something
	ne	int	ring	cause	uction	ormatio	olitical	

```
necessary
      ea      est      ture      where      change      certain      universi
hemselves
      ra      sta      cont      tatio      follow      increas      function
themselve
      ce      cti      ents      could      positi      relatio      informat
plication
      li      ica      cons      efore      govern      special      niversit
anization
      ch      ist      rati      contr      sition      process      iversity
according
      ll      ear      thin      hould      merica      against      lication
differenc
      be      ain      part      shoul      direct      problem      experien
operation
      ma      one      form      tical      bility      nstitut      structur
ifference
      si      our      ning      gener      effect      politic      determin
rganizati
      om      iti      ecti      esent      americ      ination      ollowing
organizat
      ur      rat      some      great      public      univers      followin
ganizatio
```

N-gram Counts by Word Length and Position within Word

Finally we are ready to break out the results by n-gram length, by position within word (as we did for letter counts), and also by word length. You will be able to get counts for, say, the number of times the bigram "he" appears in positions 2 through 3 of 4-letter words, for example. This is the kind of tables provided by Mayzner, but with 37 million times more data (and with a few more columns). The tables are large, so we present them in separate files; for each n-gram length from n=1 to n=9, we offer a Google Fusion Table file; you can browse the table online, or download it (with the "File > Download" menu item). We also offer all the files rolled up into a .zip file, or in a fusion table folder:

N	
1	
2	
3	
4	
5	
6	
7	
8	
9	
*	

N-gram column notation

Each column is given a name of the form "*wordlength* **/** *start* **:** *end*". For example, "4/2:3" means that the column counts the number of ngrams that occur in 4 letter words (such as "then"), and only in position 2 through 3 (such as the "he" in "then"). We aggregate counts with a notation involving a "*": the notation "*/2:3" refers to the second through third position within words of any length; "4/*" refers to any start positions in words of length 4; and "*/*" means any start position within words of any length. Finally, we also aggregate counts for positions near the ends of words: the notation "*/-3:-2" means the third-to-last through second-to-last position in words of any length (for example, this would be the bigram "he" for the words "hen", "then", "lexicographer", and "greatgrandfather").

Closing Thoughts

Technology has certainly changed. Here's where you would typically see a comparison saying that if you punched the 743 billion words one to a card and stacked them up, then assuming 100 cards per inch, the stack would be 100,000 miles high; nearly halfway to the moon. But that's silly, because the stack would topple over long before then. If I had 743 billion cards, what I would do is stack them up in a big building, like, say, the Vehicle Assembly Building (VAB) at Kennedy Space Center, which has a capacity of 3.6 million cubic meters. The cards work out to only 2.9 million cubic meters; easy peasy; room to spare. And an IBM model 84 card sorter could blast through these at a rate of 2000 cards per minute, which means it would only take 700 years per pass (but you'd need multiple passes to get the whole job done).

Aren't you glad I'm providing these tables online, rather than on cards? If you use these tables to do some interesting analysis, leave a comment to let us know. Enjoy!

Peter Norvig

View the forum thread.