

WORD—INITIAL AND WORD—FINAL NGRAM FREQUENCIES

David C. Rubin
Lawrence University

WORD—INITIAL AND WORD—FINAL NGRAM FREQUENCIES

David C. Rubin^a
Lawrence University

Abstract. Every word-initial and word-final letter cluster, or ngram, that occurred in 30 or more different words in one million words of running text is listed along with the number of different words and the total number of words it appeared in. The relation of this list to other counts is discussed.

Letters in the initial and final positions of words have a special role in reading. They begin or end not only their word, but also a phoneme, a syllable, or morpheme, and usually a spelling pattern (Fries, 1963; Gibson, Pick, Osher, and Hammond, 1962). Thus, clusters of letters in initial and final positions are more likely than medial clusters to constitute one or more of these units. Children just learning to read use initial and final letters as cues in word recognition more often than they use either medial letters or word shape (Marchbanks and Levin, 1965; Williams, Blumberg, and Williams, 1970). While adults do not use initial and final letters as often (Williams, Blumberg, and Williams, 1970), they still are more likely to perceive initial and final letters more accurately than medial letters in tachistoscopic presentation (Bruner and O'Dowd, 1958; Haslerud and Clark, 1957). Initial and final letter clusters are even important in remembering words when perceptual processing is removed: people in the tip-of-the-tongue state are more likely to remember initial and final letter clusters than medial clusters (Brown and McNeill, 1966; Rubin, 1975).

While many researchers account for the advantages of initial and final letters in terms of higher information, the initial and final letters of a word actually carry less information: that is, they are more redundant (Olivier, 1970). Thus, a list of initial and final letter clusters captures more of the redundancy of English words than lists of clusters at other positions.

This paper presents a table of initial and a table of final letter clusters for use in psychological experiments as well as in the teaching of reading. Instead of listing only bigrams or trigrams, as is commonly done, letter clusters of all

^aI wish to thank Barry Chertow and Bruce Grodnick for the programming that produced Tables I and II. Reprints may be obtained from the author, Department of Psychology, Duke University, Durham, North Carolina 27706.

lengths (i.e., ngrams) will be listed. After describing the preparation of the tables, their merits relative to other existing counts will be discussed.

The Kucera and Francis (1967) word count, which samples various categories of modern written prose, was used in preparing the tables. All words in the count that contained symbols other than the 26 letters of the alphabet and the symbols ' and — were removed. This left 999,464 of Kucera and Francis' 1,014,232 running words or tokens, and 46,342 of Kucera and Francis' 50,406 distinct words or types. For this reduced sample, every initial and final ngram that occurred in 30 or more types was listed, regardless of its length, along with the number of types and tokens it appeared in. The number of types, or distinct words, is a measure of how many different words the ngram appears in: that is, how often one would expect to find the ngram in sampling dictionary entries. It indicates the range of the ngram's use. The number of tokens is a measure of how many words in the total sample the ngram appeared in: that is, how often one would expect to find the ngram in sampling prose. It indicates the frequency of occurrence of the ngram.

The initial ngrams are in alphabetical order in Table I. The final ngrams are in alphabetical order from the last letter, instead of the first letter, in Table II. Listing the words alphabetically instead of by rank order or ngram length makes it easier to note transitional probabilities. For example, examining Table II it is easy to see that for the 3,502 types and 31,345 tokens ending in G, 3,252 types and 27,274 tokens end in ING. Thus if a word ends in G it is quite likely to end in ING.

Ngrams that did not occur in 30 or more types were not listed due to space considerations, and because the reliability of the frequencies for the less frequent ngrams is lower. For initial ngrams the 30 types limit represents only a minor restriction, as the alphabetical list of the Kucera and Francis (1967) count can be used to obtain values for less frequent ngrams. The alphabetical list of the Kucera and Francis count also provides a convenient way of finding the source words for the initial ngrams listed in Table I.

There are other counts of ngrams and of letter frequency as a function of position (e.g., Baddeley, Conrad, and Thomson, 1960; Dobby and Resnikoff, 1964; Moser, 1969; Pratt, 1942; Thorndike, 1941; Underwood and Schulz, 1960). The counts to be reviewed here have definite advantages over the current tables for certain applications. For example, as Bourne and Ford (1961) and Ohlman (1958) provide counts based on samples of subject names and proper names instead of samples of continuous text, their counts should be used in studying these domains.

Caldwell, Packham, and Nix (1974) provide a listing of all ngrams up to heptagrams with their source words. Thus, it is easy to see what words account for the use of every ngram. The size of their listing (3½ feet of bound output), however, makes the count inconvenient to use except as a data base for computer accessing.

Based on a sample of 100,000 running words Dewey (1970) presents grapheme-phoneme correspondences as a function of word position for single letters and for those bigrams that can be transcribed as a single phoneme. Both the number of types and tokens are given. Where knowledge of the ways in which single letters are usually pronounced is needed Dewey's (1970) book should be used.

Mayzner and Tresselt (1965) and Mayzner, Tresselt, and Wolin (1965a,b,c) present, as a function of word length and word position, a complete list of the number of times each ngram appears in 20,000 words of running text for ngrams up to pentagrams. Where frequency of tokens is needed as a function of position in a word or as a function of word length, this widely used count is invaluable. If only ngram frequency is needed, however, several limitations should be considered. First, no types are reported so that the number of distinct words each ngram appears in is not known. Thus, all the noted occurrences of an ngram might be due to one or two distinct words. This is also true of most other counts cited. Second, the sample size of 20,000 words is relatively small, especially if a reliable estimate of the low frequency, longer ngrams is to be obtained. The third and most serious caution is that because the authors wanted to provide information on position in a word, no words of 1, 2, or more than 7 letters were included in their sample. This is desirable if recognition of words of fixed length is being studied (e.g., Mason, 1975), however, it severely underestimates prefixes, suffixes and inflections by excluding longer words from the sample. For example, the ending TION occurs 235 times per 20,000 in the count presented here, but only 24 times per 20,000 in the Mayzner et al. count. The removal of words of 1, 2 and more than 7 letters also changes the frequency distribution and in this way the representativeness of the sample for finding ngrams (Landauer and Streeter, 1973).

Rawlinson (1976), using words of 4 or more letters in the Thorndike and Lorge (1944) word count as a sample, lists the number of tokens each bigram appears in for bigrams appearing more than one per 20,000 in an initial, final or other position. The limitations caused by removing words of certain letter length and of not providing the number of types as well as tokens have been already noted. The count provides an excellent source of bigram frequency, but only of bigram frequency, as a function of position.

Zettersten (1969), using Kucera and Francis (1967) as a sample, provides both type and token frequency for single letters and for initial and final consonant clusters. Zettersten not only provides this for the sample as a whole but also individually for each of the 15 categories of prose sampled by Kucera and Francis (e.g., Religion, Popular Lore, Science Fiction, Humor). The basic difference from the present count is that Zettersten uses consonant clusters and not ngrams: that is, Zettersten does not count vowels. Thus, he counts TH and THR as initial clusters but does not count THO or THRO. Likewise, the

Table 1: Initial ngrams, page 1

CLUSTER TYPES	TOKENS	CLUSTER TYPES	TOKENS	CLUSTER TYPES	TOKENS			
A	2,805	115,517	BLE	33	134	COND	36	558
AB	160	3,213	BLO	76	512	CONF	88	720
ABS	49	340	BLOO	31	193	CONG	35	426
AC	233	3,697	BLU	53	275	CONS	159	1,900
ACC	116	1,293	BO	428	4,252	CONST	49	508
ACCE	34	411	BDA	35	501	CONT	147	2,108
ACCO	44	589	BOL	30	84	CONTE	33	318
ACT	32	1,449	BON	41	227	CONTR	64	891
AD	243	2,556	BOO	62	453	CONTRA	35	324
ADM	40	483	BOR	39	314	CONV	64	529
ADMI	33	471	BOU	44	279	CONVE	41	346
ADV	55	583	BR	456	3,449	COO	54	481
AF	80	1,801	BRA	110	464	COR	135	944
AFF	47	400	BRE	81	520	CORR	43	293
AG	88	2,465	BREA	47	416	COS	38	534
AI	68	799	BRI	100	1,061	COU	133	4,057
AIR	38	444	BRO	119	1,204	COUN	61	1,074
AL	317	7,789	BRU	39	179	COUNT	51	916
ALL	100	3,609	BU	387	7,020	COUNTE	31	120
ALT	32	514	BUC	38	100	COUR	34	919
AM	132	2,237	BUL	42	198	CR	361	2,258
AN	364	37,664	BUR	97	504	CRA	77	333
ANA	53	312	BUS	38	616	CRE	79	634
ANG	34	346	BUT	31	4,498	CRI	55	512
ANN	37	434	BY	34	5,383	CRO	82	462
ANT	112	397	C	4,314	48,478	CROS	31	180
ANTI	68	253	CA	756	9,344	CRU	50	193
AP	180	2,350	CAL	92	1,134	CU	190	1,507
APP	127	1,997	CAM	58	1,010	CUR	69	584
APPR	50	670	CAN	90	2,593	D	2,616	30,324
AR	301	8,024	CAP	74	640	DA	229	3,129
ARC	45	206	CAR	190	1,810	DAN	49	488
ARCH	35	143	CAS	63	789	DAR	39	385
ARM	39	494	CAT	78	570	DE	870	9,231
ARR	35	382	CE	147	2,019	DEA	51	939
ART	56	767	CEN	51	1,010	DEC	102	1,040
AS	211	9,679	CENT	40	978	DEF	74	664
ASS	100	1,308	CER	31	596	DEFE	35	379
ASSE	31	268	CH	544	5,930	DEL	85	511
AST	32	96	CHA	205	2,456	DELI	37	298
AT	124	7,063	CHAN	34	803	DEM	64	612
ATT	71	1,242	CHAR	65	944	DEMO	39	358
AU	133	1,063	CHE	89	567	DEN	52	339
AUT	62	634	CHI	98	1,355	DEP	76	817
AUTO	31	272	CHO	58	495	DES	103	1,516
AV	50	675	CHR	51	465	DET	65	766
AW	33	800	CHU	32	539	DETE	43	522
B	2,839	46,281	CI	109	1,339	DEV	54	1,041
BA	527	4,568	CIR	32	309	DI	778	7,764
BAC	40	1,208	CIRAC	31	308	DIA	63	284
BAL	73	484	CL	319	3,300	DIF	31	966
BALL	32	271	CLA	104	1,089	DIFF	31	966
BAN	70	369	CLE	61	739	DIR	31	760
BAR	117	658	CLI	52	288	DIS	411	2,794
BAS	56	839	CLO	73	883	DISA	46	265
BAT	53	345	CO	1,842	22,465	DISC	88	788
BE	525	18,199	COA	40	310	DISCO	38	297
BEA	80	733	COL	149	1,537	DISP	52	314
BEE	34	2,627	COLL	61	705	DISS	32	92
BEL	75	913	COLO	41	414	DIST	66	878
BEN	51	402	COM	377	5,636	DIV	51	446
BER	63	217	COMM	127	1,919	DO	305	6,065
BES	43	558	COMME	32	307	DOU	43	348
BET	31	1,227	COMMU	36	684	DOUB	33	294
BI	195	1,655	COMP	180	2,326	DOW	35	1,041
BIR	31	237	COMPA	33	748	DR	236	2,444
BL	281	1,613	COMPL	47	751	DRA	73	587
BLA	87	549	CON	684	7,703	ORE	42	408
BLAC	37	289	CONC	86	1,118	DRI	43	627
BLACK	37	289	CONCE	53	836	ORO	33	314

Table 1: Initial ngrama, page 2

CLUSTER TYPES	TOKENS	CLUSTER TYPES	TOKENS	CLUSTER TYPES	TOKENS			
DRU	31	192	FOU	42	338	I	1,759	68,313
DU	137	1,386	FOU	54	1,244	ID	59	789
E	1,834	24,481	FR	315	7,338	IL	49	372
EA	82	2,584	FRA	91	689	ILL	42	350
EAR	39	938	FRAN	46	396	IM	248	2,310
EC	41	518	FRE	108	1,133	IMM	38	351
ED	62	857	FREE	38	511	IMP	169	1,531
EF	31	946	FRI	52	635	IMPE	38	106
EFF	31	946	FRO	48	4,778	IMPR	45	396
EI	45	625	FU	151	1,855	IN	1,145	33,770
EL	174	1,538	FUL	30	430	INA	33	128
ELE	72	858	FUN	38	585	INC	132	1,639
ELEC	46	527	FUR	33	466	INCO	43	218
ELECT	44	516	G	1,565	17,228	IND	144	1,977
ELECTR	34	326	GA	279	1,784	INDE	31	468
EM	147	1,199	GAL	53	151	INDI	66	1,017
EMB	39	148	GAR	53	290	INDU	32	456
EMP	44	556	GAS	32	173	INE	36	158
EN	341	3,994	GE	177	2,806	INF	104	834
ENC	45	265	GEN	64	1,141	INH	32	133
END	48	725	GENE	36	897	INS	164	1,667
ENG	42	724	GER	34	261	INST	62	896
ENT	64	905	GI	105	1,692	INSU	30	152
EP	48	148	GL	151	733	INT	291	4,347
EPI	36	121	GLA	43	354	INTE	213	2,229
EQ	41	556	GLO	49	190	INTER	162	1,632
EQU	38	550	GO	194	4,411	INTR	47	194
ER	64	310	GOD	30	936	INV	89	852
ES	82	1,014	GR	448	4,546	INVE	35	330
EST	30	491	GRA	195	1,244	IR	87	341
ET	30	239	GRAN	46	301	IRR	48	157
EV	106	3,483	GRE	91	1,564	IS	57	10,857
EVA	30	134	GREE	40	334	J	487	5,426
EVE	49	2,900	GRI	62	254	JA	119	876
EX	433	5,355	GRO	81	1,427	JAC	31	231
EXA	37	663	GU	156	1,035	JE	103	625
EXC	68	882	GUI	32	245	JO	117	1,580
EXE	31	274	H	1,910	54,627	JU	121	2,057
EXP	148	2,291	HA	584	16,403	K	547	5,067
EXPE	51	1,315	HAL	106	668	KA	89	288
EXPL	35	428	HALF	68	410	KE	102	1,116
EXT	71	710	HALF-	61	103	KI	130	1,245
EXTR	39	293	HAM	36	151	KIN	33	549
F	2,089	41,087	HAN	102	1,372	KN	76	1,987
FA	351	5,196	HAND	68	1,084	KNO	32	1,323
FAC	42	1,564	HAR	140	971	KO	61	167
FAI	45	592	HARD	41	486	KR	37	85
FAL	30	342	HE	401	17,884	L	1,638	23,759
FAR	47	773	HEA	117	2,096	LA	378	5,068
FAS	33	281	HEAD	35	689	LAN	58	.608
FAT	35	433	HEAR	31	788	LAT	30	846
FE	210	3,488	HEL	45	1,012	LAU	43	263
FEA	35	431	HER	82	4,269	LE	319	5,206
FEL	33	644	HI	216	12,402	LEA	79	1,793
FER	44	108	HIG	68	990	LEG	43	461
FI	351	6,427	HIGH	68	990	LI	366	7,200
FIL	48	555	HIGH-	41	80	LIB	37	321
FIN	80	1,626	HO	437	5,947	LIG	36	535
FIR	45	1,926	HOL	53	640	LIGH	30	525
FL	295	1,582	HOM	58	753	LIGHT	30	525
FLA	95	401	HOME	43	712	LIN	58	802
FLE	34	233	HON	34	291	LIT	36	1,173
FLO	87	568	HOO	36	87	LO	389	5,436
FLU	40	173	HOR	55	396	LOC	34	618
FO	406	15,108	HOU	40	1,193	LON	77	1,276
FOL	35	748	HU	177	1,622	LONG	61	1,127
FOO	38	422	HUM	41	512	LONG-	31	113
FOR	210	12,329	HUN	34	451	LOO	38	1,179
FORE	57	468	HY	87	280	LOU	30	231
FORM	38	1,068	HYP	41	127	LOW	51	418

Table 1: Initial ngrams, page 3

CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS
LU	145	621	OV	172	1,738	QUI	60	725
M	2,510	39,658	OVE	167	1,724	R	2,525	26,228
MA	826	12,926	OVER	165	1,716	RA	353	3,335
MAC	45	297	P	3,471	39,400	RAC	32	282
MAG	52	334	PA	657	6,818	RAD	41	403
MAI	35	538	PAC	36	250	RADI	35	373
MAL	60	179	PAI	32	603	RAI	44	461
MAN	144	3,429	PAL	55	319	RAN	46	583
MAR	182	1,669	PAN	51	229	RAT	53	905
MAS	65	472	PAR	196	2,542	RE	1,403	15,998
MAT	78	1,073	PARA	59	257	RE-	42	55
ME	412	6,859	PART	47	1,647	REA	112	2,785
MEA	45	1,349	PAS	62	941	REC	167	1,871
MED	44	351	PAT	90	805	RECO	67	819
MEDI	33	326	PATR	30	130	RED	74	552
MEL	34	152	PE	463	5,127	REF	85	722
MEM	33	715	PEA	35	378	REG	74	803
MEN	44	1,055	PEN	55	326	REGI	31	327
MER	55	537	PER	214	2,994	REI	46	103
MET	74	729	PERI	30	396	REL	76	1,321
MI	453	4,830	PERS	53	887	REM	74	1,156
MIC	49	216	PET	41	203	REN	42	206
MID	46	278	PH	157	1,159	REP	132	1,514
MIL	78	1,037	PHI	44	368	REPR	38	354
MILL	30	409	PHIL	40	362	RES	191	2,636
MIN	85	1,222	PHO	58	274	RESI	37	277
MIS	115	931	PI	255	1,745	RESP	32	672
MO	509	8,625	PIC	42	498	REST	41	480
MOD	46	530	PIN	43	166	RET	68	657
MOL	35	182	PL	237	3,608	REV	91	659
MON	124	1,061	PLA	143	2,914	REVE	43	293
MONO	31	75	PLAN	38	1,016	RH	31	204
MOR	81	3,072	PLAY	33	601	RI	206	2,199
MOT	56	638	PLE	34	371	RIG	33	800
MOU	41	364	PLU	35	188	RO	347	2,962
MU	245	3,219	PO	478	5,944	ROB	33	275
MUL	56	170	POL	107	1,206	ROC	37	210
MULT	39	122	POLI	38	936	ROCK	32	196
MULTI	38	116	POLY	33	102	ROO	36	654
MUS	72	1,643	POR	56	337	ROS	31	165
MY	46	1,676	PORT	38	249	ROU	54	381
N	1,046	21,677	POS	75	1,343	RU	170	1,419
NA	177	2,387	POST	38	250	RUS	34	368
NAT	53	1,335	POT	30	224	S	5,306	69,585
NE	269	6,013	PR	905	12,020	SA	406	6,451
NEU	31	111	PRA	47	478	SAL	77	571
NEW	60	2,055	PRE	326	3,501	SAN	64	376
NI	127	1,054	PRE-	33	43	SAT	33	441
NO	376	11,102	PREC	43	306	SC	344	2,648
NOM	148	366	PRED	32	118	SCA	66	296
NON-	86	117	PRES	76	1,836	SCH	102	1,113
NOR	50	793	PRI	102	1,457	SCHU	39	846
NOT	54	5,653	PRO	416	6,558	SCO	56	317
NU	81	1,058	PROC	36	647	SCR	70	359
O	1,121	71,806	PROF	43	419	SE	649	9,515
OB	99	1,177	PRDP	61	815	SEA	63	577
OBS	42	335	PROS	35	185	SEC	68	1,340
OC	55	649	PROT	44	376	SEE	40	2,267
OF	68	38,414	PROV	36	906	SEL	137	632
OFF	61	1,629	PS	54	194	SELF	106	263
OL	51	872	PSY	32	152	SELF-	98	208
ON	93	12,767	PSYC	31	149	SEM	52	86
ONE	53	3,600	PSYCH	31	149	SEMI	43	59
ONE-	43	115	PU	231	2,610	SEN	68	1,125
OP	85	1,880	PUN	33	86	SER	52	1,257
OPE	39	1,163	PUR	66	669	SEV	44	664
OR	164	5,934	Q	186	1,944	SEVE	43	663
OS	30	68	QU	185	1,942	SEVEN	30	210
OU	139	4,275	QUA	74	526	SH	486	8,457
OUT	121	2,839	QUE	38	610	SHA	134	1,267

Table 1: Initial ngrams, page 4

CLUSTER TYPES	TOKENS	CLUSTER TYPES	TOKENS	CLUSTER TYPES	TOKENS			
SHE	69	3,492	SYM	35	360	V	699	6,509
SHI	79	447	SYN	36	89	VA	155	1,636
SHO	126	2,968	T	2,421	160,282	VAL	40	612
SHOR	37	465	TA	295	3,317	VAR	32	578
SHORT	32	387	TAB	30	293	VE	184	1,691
SHR	40	121	TAL	32	556	VENM	39	134
SHU	32	135	TAN	34	150	VER	78	1,163
SI	357	4,813	TAR	36	135	VI	235	2,059
SID	34	565	TE	353	3,889	VIC	41	281
SIG	45	668	TEA	50	529	VIS	36	461
SIL	35	226	TEL	42	592	VO	107	1,020
SIM	46	742	TELE	33	226	VOL	41	342
SIN	72	1,185	TEM	33	345	W	1,379	61,516
SING	35	353	TEMP	32	344	WA	342	15,606
SIX	30	363	TEN	73	649	WAL	57	749
SK	108	583	TER	58	552	WAR	78	1,013
SKI	56	326	TERR	33	227	WAS	33	10,404
SL	191	1,265	TH	467	110,311	WAT	55	880
SLA	51	271	THA	35	12,729	WATE	36	568
SLI	41	366	THE	127	85,285	WATER	36	568
SLO	46	338	THER	52	3,320	WE	253	9,724
SM	88	1,222	THI	88	7,111	WEA	43	479
SN	102	375	THIR	37	346	WEL	86	1,195
SNA	31	171	THO	47	2,328	WELL	67	1,031
SO	399	8,980	THR	115	2,338	WELL-	54	105
SOC	47	822	THRE	61	921	WH	191	14,366
SOCI	36	797	THREE	43	696	WHE	41	3,851
SOL	73	594	THRE-	38	75	WHI	96	4,931
SOM	34	2,781	THRO	30	1,324	WHIT	39	452
SON	32	395	THU	33	418	WHO	33	3,077
SOU	60	1,160	TI	186	2,873	WI	275	13,117
SP	398	3,685	TIM	37	2,037	WIL	74	2,843
SPA	68	459	TO	302	31,193	WILL	35	2,604
SPE	113	1,578	TOP	32	286	WIN	74	769
SPEC	64	808	TOR	40	161	WIT	34	8,417
SPI	54	489	TOU	30	350	WO	218	7,436
SPL	31	104	TR	514	5,012	WOO	47	245
SPD	60	546	TRA	234	2,023	WOOD	32	211
SPR	45	402	TRAN	112	570	WOR	96	3,314
SQ	60	320	TRANS	105	552	WORK	35	1,389
SQU	59	316	TRE	63	767	WR	79	1,151
SQUA	32	246	TRI	109	821	Y	190	8,540
ST	863	12,123	TRO	48	378	YA	48	275
STA	249	4,683	TRU	57	718	YE	55	2,526
STAN	32	596	TU	128	1,156	YO	56	5,602
STAR	40	709	TUR	44	805	Z	77	215
STAT	55	2,283	TW	118	1,977			
STE	144	977	TWE	34	276			
STI	74	1,100	TWI	31	197			
STO	104	1,413	TWO	51	1,502			
STR	200	2,485	TWO-	46	82			
STRA	83	573	TY	36	473			
STRE	42	868	U	994	11,482			
STRI	37	342	UN	793	5,493			
STU	73	1,140	UNA	56	199			
SU	581	7,060	UNC	74	303			
SUB	162	1,070	UNCO	37	135			
SUBS	54	394	UND	139	1,481			
SUC	30	1,704	UNDE	107	1,389			
SUN	43	368	UNDER	92	1,350			
SUP	110	1,131	UNE	36	143			
SUPE	64	269	UNF	43	153			
SUPER	64	269	UNI	78	1,641			
SUPP	35	791	UNL	30	237			
SUR	75	1,107	UNP	41	95			
SW	153	843	UNR	44	82			
SWA	39	110	UNRE	37	74			
SWE	58	314	UNS	76	159			
SWI	43	314	UP	71	2,711			
SY	105	1,078	UR	32	226			

Table 2: Final ngrams, page 1

CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS
-	33	45	RMED	37	373	WARD	32	837
'	251	413	NED	354	3,615	ORD	57	748
S*	250	412	ENED	77	615	FORD	36	179
ES*	44	76	INED	103	1,209	E	5,580	202,651
RS*	73	122	AINED	49	696	BE	42	6,691
ERS*	59	97	ONED	69	341	CE	506	10,940
A	972	28,445	IONED	35	218	ACE	68	2,016
CA	42	308	RNED	31	828	ICE	70	2,054
OA	51	263	PED	164	1,220	NCE	312	5,968
EA	38	733	PPED	70	595	ANCE	143	2,038
IA	177	979	RED	495	4,441	ENCE	152	2,697
KA	32	91	ARED	43	703	DE	283	4,442
LA	79	360	ERED	234	1,691	ADE	73	1,610
LLA	31	81	DERED	31	398	IDE	111	1,838
MA	71	369	TERED	72	413	SIDE	36	1,017
NA	114	506	IRE	47	559	UDE	42	430
ANA	30	142	ORED	61	321	EE	135	2,572
RA	90	469	URED	59	391	REE	45	1,198
SA	33	92	SED	242	3,193	FE	36	1,182
TA	78	508	ASED	34	429	GE	296	4,598
VA	34	96	ISED	33	381	AGE	156	2,086
B	110	1,020	OSED	40	489	TAGE	33	415
C	641	5,081	SSED	74	810	DGE	40	566
IC	573	4,595	ESSED	39	392	NGE	43	753
HIC	33	110	TED	1,012	8,515	HE	53	82,530
MIC	51	530	ATED	381	2,360	IE	157	843
NIC	65	350	CATED	37	331	KE	187	3,854
ONIC	36	182	IATED	31	157	AKE	40	1,782
RIC	56	267	LATED	64	356	IKE	94	1,672
TIC	257	1,526	ULATED	31	138	LIKE	85	1,462
ATIC	56	394	NATED	40	191	-LIKE	44	46
MATIC	36	226	RATED	76	412	LE	950	11,054
ETIC	41	251	CTED	88	1,134	ALE	62	420
STIC	88	514	ECTED	47	808	BLE	447	3,813
ISTIC	62	292	ITED	55	974	ABLE	321	2,409
D	5,651	105,478	NTED	129	1,090	NABLE	33	218
AD	114	7,491	ENTED	57	347	RABLE	49	380
EAD	54	1,484	RTED	61	592	TABLE	76	596
ED	4,559	41,506	STED	91	799	IBLE	74	1,003
BED	68	527	ESTED	31	367	CLE	30	389
CED	124	1,261	TTED	52	326	OLE	34	320
ACED	42	323	UTED	36	260	ILE	82	1,330
NCED	37	457	UED	35	339	LLE	82	179
DED	281	2,519	VED	94	1,800	ILLE	61	144
ADED	49	209	WED	63	682	VILLE	51	114
NOED	99	818	OWED	41	558	OLE	33	577
ENOD	31	408	YED	69	534	PLE	34	1,768
EED	41	1,022	AYED	30	297	TLE	40	1,399
GED	156	977	ZED	179	773	ME	179	9,383
GGED	37	143	IZED	155	679	AME	39	2,222
NGED	37	292	LIZED	62	243	IME	55	1,921
HED	208	1,756	ALIZED	39	176	TINE	44	1,792
CHED	87	714	ID	118	4,266	OME	53	3,588
TCHED	31	242	AID	31	2,455	NE	468	8,020
SHED	92	882	LD	150	8,956	ANE	42	341
ISHED	47	594	ELD	52	695	INE	220	2,065
IED	162	1,682	IELD	40	402	LINE	51	516
FIED	62	370	FIELD	35	351	ONE	101	4,975
IFIED	58	323	OLD	65	1,813	PE	89	1,080
RIED	36	622	ND	299	36,401	YPE	35	253
KED	190	2,156	AND	120	31,050	TYPE	35	253
CKED	87	498	LAND	67	822	RE	427	21,800
ACKED	32	151	END	54	1,152	ARE	54	5,340
LED	423	2,843	UND	76	1,981	ERE	43	8,192
BLED	35	195	OUND	62	1,879	IRE	75	849
ELED	43	130	OD	103	2,533	ORE	75	4,094
ILED	47	397	OOD	80	1,745	URE	121	2,878
LLED	96	1,108	WOOD	37	143	TURE	72	1,564
ILLED	37	285	RD	233	3,381	SE	379	10,350
MED	147	1,333	ARD	153	2,312	ASE	43	1,131

Table 2: Final ngrams, page 2

CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS
ISE	83	829	URING	34	741	NAL	150	2,477
OSE	52	2,124	SING	170	1,332	ONAL	104	1,819
RSE	39	938	ISING	31	278	IONAL	90	1,573
USE	78	2,461	SSING	49	387	TIONAL	72	1,322
OUSE	46	696	TING	639	4,074	ATIONAL	34	829
HOUSE	37	669	ATING	212	864	RAL	82	2,185
TE	649	7,112	MATING	32	87	TAL	74	1,024
ATE	394	4,047	RATING	42	204	NTAL	35	339
LATE	52	384	CTING	58	291	ENTAL	31	317
MATE	32	236	ITING	38	387	UAL	44	1,115
NATE	58	293	NTING	71	391	EL	217	1,890
RATE	84	769	RTING	36	205	IL	103	1,633
ERATE	31	206	STING	56	448	AIL	40	396
TATE	30	945	TTING	63	650	LL	274	11,445
ITE	85	1,427	VING	101	1,279	ALL	76	4,990
TTE	36	145	IVING	32	461	ELL	101	1,837
UTE	48	457	WING	64	768	ILL	52	3,694
UE	117	1,798	OWING	39	611	OL	74	1,277
GUE	31	256	YING	131	1,163	UL	126	1,109
QUE	42	239	FYING	30	71	FUL	108	987
VE	415	10,556	ZING	87	250	M	756	16,907
AVE	33	4,771	IZING	68	159	AM	110	1,559
IVE	309	3,814	ONG	35	1,986	HAM	31	109
SIVE	67	541	RG	44	143	EM	41	3,038
TIVE	199	1,925	H	825	26,123	IM	46	3,036
ATIVE	79	723	AH	48	153	OM	100	5,692
CTIVE	55	734	CH	218	9,229	OOM	40	570
ZE	119	656	ACH	32	1,344	ROOM	33	525
IZE	89	516	ICH	30	3,738	RM	52	1,287
F	184	42,126	NCH	35	404	SM	197	739
FF	74	1,047	TCH	54	454	ISH	187	688
OFF	34	688	GH	58	2,934	LISM	55	238
LF	30	1,702	UGH	37	2,367	ALISM	48	213
G	3,502	31,345	OUGH	30	2,321	NISM	42	176
NG	3,344	30,025	SH	199	1,638	UM	162	824
ING	3,252	27,274	ISH	126	1,094	IUM	58	254
BING	34	128	TH	229	11,693	N	3,608	87,784
CING	74	461	ATH	31	547	AN	626	13,866
DING	258	2,554	RTH	37	750	EAN	30	527
ADING	36	370	I	356	6,057	IAN	181	899
LDING	39	327	LI	49	90	RIAN	39	154
NDING	83	907	NI	52	95	MAN	201	2,598
GING	117	603	RI	32	95	SMAN	30	113
NGING	41	327	TI	34	67	EN	489	14,708
HING	148	2,568	K	591	9,173	DEN	59	451
CHING	63	490	AK	44	348	EEN	41	3,952
SHING	58	288	CK	258	3,347	KEN	42	569
ISHING	31	149	ACK	80	1,740	MEN	100	1,323
KING	214	2,068	ICK	68	535	TEN	55	939
AKING	48	626	OCK	56	560	VEN	36	1,886
CKING	65	268	EK	37	499	IN	455	25,815
OKING	32	273	NK	66	1,094	AIN	71	2,279
LING	329	1,787	OK	50	1,244	EIN	34	130
ELING	30	270	OOK	39	1,226	KIN	33	134
ILING	36	220	RK	69	1,763	LIN	41	241
LLING	71	551	ORK	31	1,184	TIN	36	201
MING	107	720	L	1,739	31,924	NN	35	131
NING	248	2,667	AL	885	14,067	ON	1,663	26,541
ENING	61	468	CAL	207	2,649	ION	1,119	14,851
INING	60	633	ICAL	192	2,199	SION	128	2,013
AINING	31	441	GICAL	49	310	SSION	42	837
NNING	34	538	LOGICAL	45	294	TION	936	11,767
ONING	36	145	LOGICAL	44	293	ATION	634	6,245
PING	130	697	OLOGICAL	40	255	CATION	82	855
PPING	54	237	TICAL	47	615	ICATION	62	488
RING	317	2,508	EAL	35	697	FICATION	39	192
ARING	49	393	IAL	134	2,368	IFICATION	38	190
ERING	150	685	RIAL	34	649	IATION	34	396
TERING	46	130	TIAL	33	476	LATION	69	564
ORING	32	125	MAL	30	399	ULATION	32	273

Table 2: Final ngrams, page 3

CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS
NATION	68	638	TTER	51	1,200	IES	476	3,708
INATION	44	368	VER	91	3,938	CIES	35	272
RATION	97	1,152	OVER	33	1,630	LIES	36	307
ERATION	32	383	WER	43	868	RIES	109	942
TATION	78	643	OWER	32	633	ARIES	36	146
ZATION	75	412	IR	46	3,541	TIES	151	1,424
IZATION	75	412	OR	322	17,816	ITIES	105	958
LIZATION	39	160	TOR	137	1,121	LITIES	37	365
CTION	122	2,332	ATOR	59	322	KES	44	477
ECTION	45	890	CTOR	35	518	LES	241	1,896
ITION	65	1,179	UR	87	3,092	BLES	47	216
PTION	36	378	DUR	48	2,858	MES	54	1,337
RON	39	153		\$11,021	128,971	NES	138	981
SON	136	1,545	S	1,817	5,694	INES	67	537
TON	151	1,006	A'S	78	189	ONES	30	292
RN	104	1,848	O'S	91	322	OES	30	771
ERN	35	870	RD'S	30	68	PES	33	269
ORN	41	314	E'S	267	915	RES	153	1,212
UN	51	654	LE'S	34	87	IRES	30	173
WN	111	2,995	NE'S	47	148	URES	59	710
OWN	94	2,858	G'S	37	95	TURES	39	405
O	574	42,364	H'S	43	120	SES	237	1,906
CO	46	258	I'S	39	71	ASES	30	399
DO	56	1,501	K'S	48	108	ISES	32	183
GO	35	1,107	L'S	87	192	SSES	71	522
IO	59	332	M'S	52	111	ESSES	43	261
LO	42	158	N'S	303	905	USES	40	276
NO	44	2,325	AN'S	77	290	TES	243	1,901
RO	45	351	MAN'S	42	213	ATES	151	1,278
TO	79	28,286	EN'S	33	118	RATES	30	193
P	386	6,154	IN'S	41	78	ITES	31	170
AP	35	263	ON'S	117	291	UES	51	601
EP	39	754	SON'S	35	98	VES	121	1,454
IP	93	1,005	TON'S	31	55	IVES	54	540
HIP	64	707	O'S	53	129	ZES	41	93
SHIP	61	661	R'S	263	692	FS	34	103
MP	41	242	ER'S	203	513	GS	277	1,839
OP	63	783	TER'S	41	82	NGS	222	1,499
UP	77	2,513	OR'S	41	106	INGS	205	1,363
-UP	35	72	S'S	35	74	HS	67	425
R	2,374	59,096	T'S	152	999	THS	37	341
AR	258	4,735	NT'S	30	147	IS	221	23,645
EAR	60	2,021	Y'S	180	569	SIS	65	639
LAR	84	908	EY'S	40	77	KS	201	1,509
ULAR	48	571	AS	159	20,765	CKS	79	364
ER	1,630	28,662	BS	58	282	NKS	30	246
BER	48	1,237	CS	100	669	LS	354	2,943
DER	158	2,316	ICS	87	634	ALS	127	960
LDER	30	284	TICS	41	340	ELS	79	505
NDER	57	1,065	DS	345	4,129	ILS	41	261
FER	33	294	ADS	38	263	LLS	70	685
GER	98	1,050	IDS	34	173	MS	173	2,385
NGER	37	572	NDS	104	1,679	OMS	30	186
HER	115	8,192	ANOS	38	631	NS	986	7,868
CHER	30	193	RDS	81	962	ANS	166	1,269
THER	54	4,720	ARDS	58	475	IANS	85	345
IER	89	488	ES	2,481	22,355	ENS	80	435
KER	89	410	CES	158	1,974	INS	109	715
LER	121	508	NCES	80	740	AINS	39	396
LLER	35	222	ANCES	43	359	ONS	551	4,789
MER	63	558	ENCES	32	359	IONS	455	4,132
NER	97	848	DES	117	700	SIONS	67	524
PER	74	1,048	ADES	33	148	TIONS	365	3,360
RER	32	154	IDES	39	361	ATIONS	215	1,715
SEP	35	176	EES	51	493	RATIONS	37	282
TER	331	5,687	GES	98	923	TATIONS	36	252
ATER	33	1,167	AGES	55	494	CTIONS	57	634
ETER	37	196	HES	137	853	ITIONS	32	373
NTER	34	542	CHES	82	566	RNS	33	261
STER	82	601	SHES	45	158	OS	102	349

Table 2: Final ngrams, page 4

CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS	CLUSTER	TYPES	TOKENS
PS	176	1,456	FT	44	901	BLY	102	847
IPS	44	304	HT	128	4,404	ABLY	67	648
OPS	36	210	GHT	120	4,389	DLY	109	687
RS	1,115	9,161	IGHT	97	3,216	EDLY	56	161
ARS	64	1,651	LIGHT	30	560	ELY	211	2,359
ERS	812	5,230	IT	144	10,692	TELY	52	673
DEERS	91	557	LT	52	1,241	ATELY	36	473
NDERS	37	84	NT	798	13,131	VELY	72	432
GERS	43	236	ANT	187	2,254	IVELY	68	376
HERS	54	620	TANT	33	700	TIVELY	40	261
KERS	63	242	ENT	503	9,353	GLY	118	361
LEERS	53	146	CENT	32	586	NGLY	114	334
NERS	66	319	DENT	36	907	INGLY	112	296
PERS	47	175	IENT	34	399	ILY	83	1,061
TERS	155	1,037	MENT	238	4,189	LLY	409	3,511
STERS	35	143	EMENT	93	1,241	ALLY	308	2,921
VERS	41	177	INT	44	705	CALLY	145	584
ORS	173	1,129	UNT	32	469	ICALLY	141	566
TORS	94	603	OT	110	6,307	TICALLY	61	250
ATORS	35	149	OOT	30	192	IALLY	32	394
URS	37	331	PT	53	954	NALLY	42	416
SS	617	5,505	RT	149	3,259	ONALLY	31	178
ASS	50	730	ART	41	1,264	ULLY	61	396
ESS	526	3,836	ERT	42	374	FULLY	56	382
LESS	128	1,060	ORT	45	1,193	NLY	33	2,221
ELESS	32	181	ST	546	12,551	RLY	58	1,221
NESS	336	1,462	AST	46	2,051	SLY	135	633
DNES	36	99	EST	210	2,593	SSLY	30	64
ENES	68	192	TEST	32	88	USLY	102	561
INES	48	511	IST	195	1,165	OUSLY	102	561
LNES	34	93	LIST	41	284	IOUSLY	50	355
SNES	37	104	ALIST	32	112	TLY	168	1,942
TNES	36	125	NIST	37	178	NTLY	97	945
RESS	31	761	OST	35	2,209	ANTLY	33	158
TS	1,023	10,789	UST	31	2,201	ENTLY	61	765
ATS	53	301	TT	74	198	MY	48	1,810
CTS	62	990	ETT	33	70	NY	97	3,436
ETS	107	712	UT	107	10,046	ONY	32	282
HTS	31	341	OUT	58	4,783	OY	39	501
GHTS	30	338	U	104	3,761	PY	41	278
ITS	74	2,411	V	50	272	RY	511	7,489
NTS	308	3,071	M	198	9,156	ARY	142	2,120
ANTS	73	494	AW	42	869	NARY	34	345
ENTS	198	2,207	EW	46	3,323	TARY	34	601
MENTS	123	1,242	OW	104	4,873	ERY	100	1,925
EMENTS	43	536	LOW	37	893	ORY	107	1,225
OTS	38	235	X	122	1,278	TORY	77	858
RTS	50	831	EX	30	380	ATORY	38	138
STS	198	1,245	Y	3,875	59,030	RRY	34	414
ISTS	130	555	AY	188	6,455	TRY	48	906
LISTS	30	97	DAY	41	1,600	URY	37	475
UTS	35	156	WAY	64	1,680	SY	53	401
US	517	4,195	BY	46	5,614	TY	586	6,454
OUS	325	2,275	CY	110	1,006	ITY	419	4,008
IOUS	135	1,227	ACY	30	219	CITY	41	664
CIOUS	46	215	NCY	65	445	LITY	179	1,154
ROUS	44	289	ENCY	41	353	ALITY	78	488
MS	79	947	DY	101	1,969	ILITY	93	647
ONS	43	542	EY	204	4,994	BILITY	78	557
YS	115	1,804	LEY	82	335	ABILITY	47	271
AYS	67	1,466	NEY	31	505	NITY	32	613
WAYS	33	656	FY	49	212	RITY	52	543
T	2,739	94,374	IFY	36	174	Z	88	288
AT	109	20,047	GY	75	447	TZ	32	70
EAT	34	1,219	OGY	45	254			
CT	119	3,089	LOGY	43	251			
ECT	51	1,438	OLOGY	40	229			
ET	247	4,741	HY	77	873			
KET	36	376	KY	71	271			
LET	36	563	LY	1,528	16,385			

occurrence of *THR* is not counted under the occurrence of *TH* because *THR* is not an instance of *TH* followed by a vowel. When consonant clusters and not ngrams are of interest Zettersten is clearly the best count to use.

REFERENCES

- BADDELEY, A.D., CONRAD, R. and THOMSON, W.E. Letter structure of the English language. *Nature*, 1960, 186, 414-416.
- BOURNE, C.P. and FORD, D.F. A study of the statistics of letters in English words. *Information and Control*, 1961, 4, 48-67.
- BROWN, R. and McNEILL, D. The "tip-of-the-tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 1966, 5, 325-337.
- BRUNER, J.S. and O'DOWD, D. A note on the informativeness of parts of words. *Language and Speech*, 1958, 1, 98-101.
- CALDWELL, E.C., PECKHAM, P.D. and NIX, D.H. Ngram frequency counts. *Developmental Psychology*, 1973, 9, 266-267.
- DEWEY, G. *Relative frequency of English spellings*. New York: Teachers College Press, 1964.
- DOLBY, J.L. and RESNIKOFF, H.L. On the structure of written English words. *Language*, 1964, 40, 167-196.
- FRIES, C.C. *Linguistics and reading*. New York: Holt, Rinehart, and Winston, Inc., 1963.
- GIBSON, E.J., PICK, A.D., OSSER, H., HAMMOND, M. The role of grapheme-phoneme correspondence in the perception of words. *American Journal of Psychology*, 1962, 75, 554-570.
- HASLERUD, G.M. and CLARK, R.E. On the redintegrative perception of words. *American Journal of Psychology*, 1957, 70, 97-101.
- KUCERA, H. and FRANCIS, W.N. *Computational analysis of present-day American English*. Providence, Rhode Island: Brown University Press, 1967.
- LANDAUER, T.K. and STREETER, L.A. Structural differences between common and rare words: Failure of the equivalence assumption for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 119-131.
- MARCHBANKS, G. and LEVIN, H. Cues by which children recognize words. *Journal of Educational Psychology*, 1965, 56, 57-61.
- MASON, M. Reading ability and letter search time: Effects of orthographic structure defined by single-letter positional frequency. *Journal of Experimental Psychology: General*, 1975, 1, 146-166.
- MAYZNER, M.S. and TRESSELT, M.E. Table of single-letter and diagram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements*, 1965, 1, (2).
- MAYZNER, M.S. TRESSELT, M.E. and WOLIN, B.R. Tables of trigram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements*, 1965a, 1, (3).
- MAYZNER, M.S., TRESSELT, M.E. and WOLIN, B.R. Tables of tetragram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements*, 1965b, 1, (4).
- MAYZNER, M.S., TRESSELT, M.E. and WOLIN, B.R. Tables of pentagram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements*, 1965c, 1, (5).
- MOSER, H.M. *One-syllable words*. Columbus, Ohio: Charles E. Merrill Publishing Company, 1969.