Phonics: A Large Phoneme-Grapheme Frequency Count Revised

Edward Fry
Rutgers University

This study is a summary and simplification of a very large phoneme-grapheme frequency count done by Hanna et al. (1966). Although the results and data from the original study have implications for teaching phonics and spelling, they were presented in a complicated and unwieldy manner. Moreover, the original study is out of print. This study, then, presents a succinct and simplified summary of the Hanna et al. results for researchers and teachers of reading and spelling.

Although phonics has been identified as an essential element of successful literacy instruction in the elementary grades (National Reading Panel, 2000), details about the nature and content of effective phonics (and spelling) programs have not been fully articulated. Questions regarding the content and sequencing of phonics instruction still exist. One approach to questions regarding the content and sequencing of phonics instruction is to examine the phoneme-grapheme content of words used in instructional contexts. Hanna, Hanna, Hodges, and Rudorf (1966) conducted such a study, examining and counting every phoneme-grapheme correspondence in a 17,310 word vocabulary (Thorndike & Lorge, 1944). The current interest in phonics has made this nearly raw basic research data very relevant to today's teachers, researchers and curriculum developers in the areas of reading, spelling, and linguistics. Unfortunately, these valuable data are reported in a U.S. Office of Education document spanning 1,716 pages that has long been out of print. Another complication is the unique coding system used by the researchers, which focused on such factors as separate counts of the phonemes in stressed and unstressed syllables, and the location of the phoneme in each syllable.

The present study has reanalyzed those basic data to simplify the original report and make it more useable. It addresses the following questions:

What are the most useful (highest frequency) phoneme-grapheme correspondences?

What are the most frequent ways of spelling those phonemes?

Answers to these questions could lead to better phonics and spelling instruction and could improve the phonics content of both commercial and teacher-made curriculum materials for reading and spelling instruction.

A similar set of questions was posed and answered (Fry, 1964) based on an earlier Stanford phoneme-grapheme research project (Moore, 1951) that used a 3,000-word count. The Hanna et al. (1966) report used a much larger (17,000+) word corpus and a much more sophisticated coding system.

Earlier work in this area has focused on determining spelling generalizations. For example, a recent study by Abbott (2000) used the large Hanna et al. count to identify reliable spelling generalizations. Her spelling generalizations were based on Clymer (1963). Another generalization study by Johnston (2001) used her own 3000-word corpus plus data from Burmeister (1968) and Clymer. The research reported here is not based on generalizations, but rather it is a strict phoneme-grapheme correspondence count.

The Hanna et al. study was one of the largest studies funded by the U.S. Office of Education up to that time. The corpus of words used in their study consisted of 17,310 different words selected from the Thorndike-Lorge *Teacher's Word Book of 30,000 Words* (1944). Hanna et al. omitted foreign words, trade names, slang, and rare words. They placed each phoneme into a 22 vowel and 30 consonant classification system according to the pronunciation given in a Mirriam-Webster dictionary.

Vowel Classification

Hanna et al. began with the Merriam Webster dictionary's vowel classification system involving 33 vowel sounds. They soon found it unworkable and simplified it to a 22 vowel system to facilitate their computerized algorithms. One of the original goals of their study was to determine how well a computer with many algorithms (rules and phoneme-grapheme information) could correctly spell each word, given the dictionary pronunciation guide. The short answer is about 50%.

The present study has a different purpose—to provide teachers and curriculum developers with usable and scientifically-based information for developing phonics and spelling instructional programs for beginning or remedial readers and spellers,

presented in a comprehensible manner. It is a compromise between the tens of thousands of little facts in the giant Hanna et al. study and the realities of the classroom. After examining the data in this study, a teacher or curriculum developer should be able to more precisely determine phonics information that is rather useless and other information that is valuable. At either end of the phoneme-use spectrum, there could be causes for disapproval; linguists and phoneticians may find this study too simple, while some classroom teachers may find it too complicated. But it does unearth and bring to light some basic data that has long been buried in a difficult scientific report (Hanna et al., 1966).

The present study further simplifies the Hanna et al. vowel classification system to 17 categories or classifications as described below and seen in the tables. This was done to make the system more comprehensible and usable for teachers.

The simplification process was complicated. For example, many dictionaries do not recognize a Long U; rather, a Long OO sound, as in "moon" or "rule," is specified. In the present study I have combined Hanna et al.'s categories U1 and O6. Teachers can call it Long U or Long OO, whichever suits them.

The Short U, as in "up," is also problematic. Phonetically it is similar to a schwa, as in the A in "ago." Technically a schwa must be in an unaccented syllable, but I have combined the Short U and Schwa (Hanna et al.'s categories U3 and Schwa) because for all practical purposes, and certainly for beginning readers and spellers, they sound the same.

The letter or grapheme R and the phoneme /r/ cause a lot of vowel difficulty. When the letter A is followed by an R there are two different phonemes for the A: the phoneme /ä/ as in "far" and the phoneme /â/ as in "vary." I have chosen to keep these in two different categories.

The letter O followed by an R gives the O a broad sound /ô/ as in "for." A few other graphemes produce /ô/ and Hanna et al. separate them into two categories, O2 & O5. I have consolidated them in the category Broad O.

The letter R modifies the Schwa or Short U sound when R follows an E, I, or U as in "her," "fur," "sir" to yield separate categories in the Hanna et al. study. The present study has all these uses, which is very much like the /r/ in "red," listed in the Short U+R vowel category of Table 1.

To summarize, the major changes between the classification system used in the present study and the Hanna et al. study involve consolidating U1 & O6, U3 & Schwa, and O2 & O5. These changes are reflected in Table 1.

Table 1. Vowel Phoneme-Grapheme Correspondences

Phoneme	Grapheme	Example	Rule	Frequency
A short	a	at	Closed syllable	4192
(A3) /a/	a-e	dance	unusual	147
A long	a	agent	Open syllable	1002
(A1) /ā/	a-e	ate	Final E	790
	ai	rain	AI digraph	208
	ay	day	AY digraph	131
	eigh	eight	rare	18
	e	cafe	rare	16
	ea	break	rare	14
	ei	veil	rare	14
	ey	they	rare	14
AR	a(r)	arm	AR digraph	474
(A5) /ä/	a	father	unusual	44
	ar-e	large	unusual	31
	ea(r)	heart	rare	18
AR	ar	vary	unusual	64
(A2) /â/	are	care	unusual	50
	air	fair	unusual	46
	ere	there	unusual	31
	ear	bear	rare	13
E short	e	end	Closed syllable	3316
(E3) /e/	ea	head	unusual	139
	e-e	fence	unusual	79
E long	e	me	Open syllable	1765
(E1) /ē/	y	very	Final Y	1801
	ee	keep	EE digraph	249
	ea	eat	EA digraph	245
	e-e	these	Final E	62
	ie	field	unusual	62
	i-e	police	unusual	44
	ey	money	unusual	40
	i	unique	unusual	38
	ea-e	peace	unusual	30
	ie-e	piece	rare	23
	ei	ceiling	rare	16
I short	i	in	Closed syllable	5346
(I3) /i/	i-e	give	unusual	339
	a-e	village	unusual	187
	y	system	unusual	100

Phoneme	Grapheme	Example	Rule	Frequency
	ui	build	rare	16
	ai	captain	rare	15
	ei	foreign	rare	11
I long	i-e	ice	Final E	555
(I1) /ī/	i	item	Open syllable	554
	y	my	Final Y	211
	igh	fight	unusual	88
	ie	pie	unusual	26
	y-e	type	rare	23
O short	o	not	Closed syllable	1558
(O3) /o/	a	what	unusual	80
	o-e	lodge	rare	20
O long	o	open	Open syllable	1876
(O1) /ō/	o-e	home	Final E	370
	oa	oat	OA digraph	126
	ow	own	OW digraph	124
	ou	mould	unusual	29
	oe	toe	rare	13
	ou-e	course	rare	10
O broad	o(r)	for	R modified	312
(O2 & O5)	a	all, war	A before L & R	165
/ô/	au	auto	AU digraph	146
	o	off, dog	unusual	123
	aw	awful	AW digraph	75
	o-e	horse	rare	17
	ough	bought	rare	15
	augh	caught	rare	12
OO short	u	pull	unusual	200
(O7) /oo/	oo	look	OO digraph — short	114
	o	woman	rare	17
	u-e	sure	rare	11
OI diphthong	oi	oil	OI digraph	92
(OI) /oi/	oy	toy	OY digraph	48
OU diphthong	ou	out	OU digraph	227
(OU) /ou/	ow	owl	OW digraph	119
U short	o	other	important schwa	1723
and schwa	u	up	closed syllable	1509
(U3 & swa)	a	ago	important schwa	1438
/u/ & /ə/	i	animal	important schwa	1347

Phoneme	Grapheme	Example	Rule	Frequency
	e	effect	important schwa	763
	ou	double	unusual	366
	o	son	unusual	112
	e-e	violence	unusual	101
	o-e	come	unusual	47
	u-e	huge	unusual	46
	o	do	unusual	37
	y	oxygen	rare	23
	ie	patient	rare	22
	eo	pigeon	rare	10
	oo-e	loose	rare	12
U long	u	unit	Open syllable	907
OO long	u-e	tune	Final E	290
(U1 & O6)	oo	moon	OO digraph—long	173
/ū / & /ōo/	ew	new	EW digraph	60
	ou	you	unusual	29
	eu	neutral	unusual	28
	ue	blue	UE digraph	27
	oo-e	goose	rare	12
	o-e	move	rare	12
Schwa R &	er	her	regular ER	1979
Short U + R	or	labor	regular OR	321
(U2 & E5)	ur	turn	regular UR	234
/ə/ & /u/	ar	dollar	regular AR	168
	ir	girl	regular IR	104
	er-e	nerve	unusual	41
	ear	earth	unusual	29
	our	journey	rare	21

Note. The symbols inside the parentheses, for example (E3), indicate Hanna's classification for that phoneme.

Correspondences with frequencies below 10 were omitted.

Consonant Categories

Although consonants are far less problematic than vowels, they are not totally free of problems. Basically each consonant (grapheme) letter represents one phoneme. There are, however, some exceptions. For example, three consonant letters represent multiple phonemes: (a) Letter X as in "box" represents the /ks/ sound, (b) Letter Q (which never appears without a U) represents the /kw/ sound as in "quick," and (c) Letter C represents two sounds—the /s/ sound as in "city" and the /k/ sound as in "cat."

Consonant digraphs—CH, SH, TH voiced, as in "this," and TH voiceless, as in "thin" each represent distinct phonemes (and should have been separate letters). The consonant digraph WH represents three phonemes; /h/, /w/, and /hw/. For example, the common word "what" can be correctly pronounced as /wot/ or /hwot/, but "why" must begin with /hw/ and "who" must begin with /h/.

A few consonant phonemes are not spelled with the expected graphemes. For example,the /j/ sound is more commonly spelled with a G as in "gem" rather than the expected J as in "just." (See Table 2 for a full presentation of consonants.)

Table 2. Consonant Phoneme-Grapheme Correspondences

Phoneme	Grapheme	Example	Rule	Frequency
B	b	boy	regular	2242
	bb	rabbit	unusual	63
C	No C phoneme, see K and S			
CH	ch	child	regular	313
	t	picture	unusual	175
	tch	catch	unusual	61
	ti	question	rare	13
D	d	dog	regular	3611
	dd	add	unusual	74
F	f	fox	regular	1580
	ph	phone	unusual	242
	ff	cliff	unusual	177
G	g	girl	regular	1178
	gg	egg	unusual	67
	gue	league	rare	21
	gu	guard	rare	19
	gh	ghost	rare	10
H	h	hot	regular	762
J	g	gem	most common	647
	j	jump	regular	218
	dge	lodge	unusual	51
	d	educate	unusual	32
	gi	legion	rare	14
K	c	cat	regular	3452
	k	kind	regular	601
	ck	back	unusual	290
	ch	chord	unusual	142

Phoneme	Grapheme	Example	Rule	Frequency
	cc	account	unusual	76
	que	unique	rare	19
/ks/	x	fox	regular for X	245
/ks/	cs	physics	unusual	26
/kz/	x	exam	unique prefix	43
/kw/	qu	quick	regular	191
L	l	like	regular	4894
	le	able	common ending	620
	ll	bell	common ending	489
	el	novel	rare	19
M	m	man	regular	3302
	mm	commute	unusual	140
	mb	lamb	unusual	27
	lm	calm	rare	17
N	n	nose	regular	7452
	en	dozen	unusual	128
	nn	funny	unusual	127
	kn	knife	silent letter	41
	on	lesson	unusual	41
	gn	sign	silent letter	32
NG	ng	sing	regular	362
	n	think	unusual	251
P	p	pig regular		3296
	pp	happy	unusual	153
Q No Q phoneme see K				
R	r	rat	regular	9134
	rr	carry	unusual	207
	wr	write	silent letter	48
	rh	rhythm	rare	16
S	s	sat	regular	4599
	c	city	before A, O, & U	1067
	ss	pass	unusual	442
	ps	psychology	rare	19
SH	sh	shoe	regular	398
	tion	action	unique suffix	820
	ci	special	unusual	119
	ssi	mission	unique suffix	51
	si	tension	unique suffix	38

Phoneme	Grapheme	Example	Rule	Frequency
	ch	chef	unique	34
	ti	patient	unique	30
	s	sugar	rare	20
T	t	toy	regular	7528
	tt	cotton	unusual	216
	ed	hooked	unique suffix	28
	bt	doubt	silent letter	11
TH voiceless	th	thank	regular	411
TH voiced	th	they	less common	149
V	v	very	regular	1485
W	w	with	regular	578
	u	suite	unusual	47
WH /hw/	wh	when	regular	89
X No X phoneme, see K				
Y	y	yard	regular	53
	i	onion	unusual	66
Z	z	zero	regular	229
	s	his	before I, E, & Y	640
	es	wives	unique suffix	44
	zz	jazz	rare	23
	ss	dessert	rare	13
ZH	si	erosion	unique suffix	55
	s	measure	unusual	34
	g	massage	rare	15

The real work of this study involved producing the tables. They simplify and summarize hundreds of pages of data from the Hanna et al. study and answer basic questions about the significance of phonics content.

Tables 1 and 2 provide all the common spellings (graphemes) for all the phonemes. The frequencies represent how often each phoneme is spelled by a particular grapheme in a 17,310-word corpus. Frequencies less than 10 are omitted; these might properly be considered exceptions as they occur less than 0.006% of the time.

I have tried to give the vowel phonemes popular names to make the information more usable for teachers. The symbols within parentheses, however, are the Hanna et al. categories. The symbols between slash marks are those often used by dictionaries.

If a majority of words follow common rule, it is listed in the rule column in Table 1. Where no rule is apparent or specified, the correspondence is classified as:

Regular—usually the most common use

Unusual—frequency less than regular but more than 50

Rare—frequency less than 50 but 10 or more

Although this classification is somewhat arbitrary, users who wish to know the frequency of any of the correspondences to help them make instructional decisions will find that the data in Tables 1 and 2 provide an empirical frequency summary of phoneme-grapheme correspondences.

Table 3. Vowels Ranked by Phoneme Frequency

Phoneme	Example/Frequency		Common alternate graphemes
I Short	in	5346	i-e give 339, a-e village 187, y system 100
A Short	at	4192	a-e dance 147
E Short	end	3316	ea head 139, e-e fence 79
Schwa R	her	1979	or labor 321, ur turn 234, ar dollar 168, ir girl 104
O Long	open	1876	o-e home 370, oa oat 126, ow own 124
E Long	me	1765	y funny 1801 , ee keep 249, ea eat 245,
U Short/ə/	other	1723	u up 1509, a ago 1438, i animal 1347, e effect 763, ou double 366, o son 112, e-e violence 101
O Short	not	1558	
A Long	agent	1002	a-e ate 790, ai rain 208,ay day 131, a danger 100
U Long	unit	907	u-e tune 290, oo moon 173, ew new 60
I Long	ice	555	i item 554, y my 211
A Broad/ä/	are	474	
O Broad/ô/	for	312	al all 165, au auto146, o off123, aw awful75
OU	out	227	ow owl 119
OO Short	pull	200	oo look 114
OI	oil	92	oy toy 48
AR /â/	vary	64	are care 50, air fair 46

Note. The frequency column is for the most common phoneme-grapheme correspondence in the 17,310 word vocabulary.

Table 4. Consonants Ranked by Grapheme Frequency

Grapheme	Frequency	Irregularities
R	9114	/r/ spelled WR write 48
T	7528	/t/ spelled ED hooked 28
N	7452	/n/ spelled EN dozen 128, KN knife 41
S	4599	S=/s/ sat 4599, S=/z/ his 640
L	4894	/l/ spelled LE able 620
C	4522	no phoneme, C=/k/ 3454 C=/s/ 1068
D	3611	
P	3296	
M	3304	
B	2242	
F	1580	/f/ spelled PH phone 242
V	1485	
G	1178	
H	762	
K	601	/k/ most commonly spelled C cat 3454, CK 290
W	578	
TH	411	/voiceless/ 411, /voiced /149
SH	389	/sh/ most commonly spelled TI action 820, CI 119
NG	362	/ng/ spelled N think 251
CH	313	/ch/ spelled T picture 175
X	245	no phoneme, X represents /ks/ box 245
Z	229	/z/ most commonly spelled S his 640
J	229	/j/ most commonly spelled G gem 647,
QU	191	no phoneme, QU represents /kw/ quit 191
WH	89	often /hw/ blend who
Y	53	Y is most commonly a vowel as in very 1801

Note. The frequencies in the frequency column are for graphemes. They are the most common or regular phoneme-grapheme correspondences in the 17,310-word vocabulary.

The graphemes and phonemes are the same for all consonants except where noted in the Irregularities column; for example, the grapheme X makes the /ks/ sound. Also noted in this column are the few instances where the phoneme is more commonly represented by another grapheme; for example, the /k/ sound is most commonly spelled by a C as in "cat." Some other relatively common spellings (graphemes) for the phoneme are also in the Irregularities column. Other less common alternate spellings are in Table 2.

Tables 3 and 4 are a further simplification of the Hanna et al. study. They answer the question: "What are the most useful (highest frequency) phoneme-grapheme correspondences?" Table 3 ranks the most common vowel phoneme-grapheme correspondences and gives a few common alternate spellings (less common correspondences). Table 4 ranks the consonants. It differs from Table 3 in that it ranks the consonants by grapheme, not phoneme. In a majority of the correspondences the phoneme and the grapheme are the same, however.

Discussion

This study validates much that is common on phonics instruction, such as the teaching of short vowels before long vowels, the final E rule, the open syllable rule, schwa, and R-modified vowels (see Table 3). The basic correspondences for consonants are important but there are some important modifications and exceptions (see Table 4).

The many ways different phonemes can be spelled, as shown in Tables 1 and 2, provide content for spelling and reading programs. And since one must limit the amount of content, particularly for beginners, the information in the tables suggests what might be eliminated and what could be emphasized.

This study does not support the teaching of phonics or phoneme-grapheme correspondences arranged in alphabetical order. The teaching of /b/ is certainly not more important than the teaching of /r/ or /t/.

The findings of this study, particularly Tables 3 and 4, which show the ranking for vowels and consonants, can be used as a checklist or tool for evaluating published reading and spelling materials. Such a checklist could help district and state curriculum coordinators in developing language arts curricula. It may also help college instructors in developing reading teacher education curricula. It could also assist teachers of English-language learners.

There are about as many ways to teach spelling as there are to teach reading. Phonics is only one way; more often it is a part of a broader collection of techniques and content. The results from this study may inform such curricula. For example, this study might help teachers in selecting categories for word sorts (Bear, Invernizzi, Templeton, & Johnston 2000).

Phonemic awareness has come to prominence partly because of the National Reading Panel's meta-analysis (Ehri et al., 2001) and the summary by Smith, Simmons, and Kameenui (1995). This study provides some content for phonemic awareness instruction.

The National Reading Panel's (2000) meta-analysis for phonics showed some benefit for systematic phonics instruction. The data in this study provides much content for a systematic or an incidental phonics program and lesson, such as the Making Words technique (Cunningham & Cunningham, 1992).

Several limitations to this study need to be identified. First, although this study presents all major phoneme-grapheme correspondences, it is not a presentation of every possible phoneme-grapheme correspondence for either reading or spelling.

The present study is a count of all the phonemes used in over 17,000 different words without regard to word frequency. For example, the digraph TH occurs in some very high frequency words like "the," "this," and "that," yet the TH digraph occurs in only 411 different words. One could argue that TH is thus more important than the 411 frequency indicates. But to weight every one of the 17,000 different words by frequency of occurrence is well beyond the scope of this study. Teachers usually solve this problem by mixing the teaching of phonics with the teaching of high-frequency sight words (e.g. Instant Words (Fry, 1999)). Perhaps some well-funded future researcher would like to work on the weighting (type/token) problem.

The present study does not deal with morphemes or meaning units (e.g., the prefix UN- has a very high frequency) rimes (phonograms), or other common letter clusters. As readers mature they tend to use larger clusters of letters than just graphemes (Adams, 1990).

One could question that the word database published in 1944 (Thorndike & Lorge) is a bit dated. Nevertheless, new words tend to have a lower frequency than the more common structure words such as "is," or common base words like "run." Thus, most of the words in the Thorndike list are still very relevant. Language does change, but it changes slowly and neologisms tend not to appear in phonics or elementary spelling lessons. However, even new words tend to use the same more common correspondences; hence, it is unlikely that using a newer English word list would substantially change the rank order of correspondences reported here.

Finally, the present study is not a child development study. It does not address which correspondences a beginning reader typically does or should learn first. However, there is an implication in the study that the more common or higher frequency correspondences should be taught first.

Phonics and phonemic awareness are keys to successful literacy acquisition (National Reading Panel, 2000). The results of this study provide reading instructors and curriculum developers with practical information for improving the precision and effectiveness of instruction in these areas.

References

Abbott, M. (2000). Identifying reliable generalizations for spelling words: The importance of multilevel analysis. *The Elementary School Journal, 102,* 233-245.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Bear, D. R., Invernizzi, M., Templeton, S., & Johnston, F. (2000). *Words their way.* Upper Saddle River, NJ: Merrill.

Burmeister, L. E. (1968). Usefulness of phonic generalizations. *The Reading Teacher, 21,* 349-356.

Cunningham, P. M., & Cunningham, J. W. (1992). Making words: Enhancing the invented spelling-decoding connection. *The Reading Teacher, 46,* 106-107.

Clymer, T. (1963) The utility of phonics generalizations in the primary grades. *The Reading Teacher, 16,* 252-258.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36,* 250-289.

Fry, E. B. (1964). A frequency approach to phonics. *Elementary English, 22,* 759-766.

Fry, E. B. (1999). *1000 Instant Words: The most common words for teaching reading, writing, and spelling.* Westminster, CA: Teacher Created Materials

Hanna, P. R., Hanna, J. S., Hodges, R. E., & Rudorf, E. H. (1966). *Phoneme-grapheme correspondences as cues to spelling improvement.* Washington, DC: U.S. Department of Health, Education, and Welfare.

Johnston, F. P. (2001). The utility of phonic generalizations: Let's take another look. *The Reading Teacher, 55,* 132-142.

Moore, J. T. (1951). *Phonetic elements appearing in a 3000 word spelling vocabulary.* Unpublished dissertation, Stanford University.

National Reading Panel. (2000). *Report of the National Reading Panel; Reports of the subgroups.* Washington, DC: National Institute of Child Health and Human Development Clearinghouse.

Smith, S. B., Simmons, D.C., & Kameenui, E. J. (1995). Synthesis of research on phonological awareness principles and implications for reading acquisition. Technical Report 21, National Center to Improve the Tools of Education, University of Oregon (A review of 29 references).

Thorndike, E. L., & Lorge, I. (1944). *The teacher's book of 30,000 words.* New York: Teachers College Press.