METHODS & DESIGNS

Empirically derived probabilities for grapheme-to-phoneme correspondences in English

RITA SLOAN BERNDT, JAMES A. REGGIA, and CHARLOTTE C. MITCHUM
University of Maryland Medical School, Baltimore, Maryland

Prior probabilities of graphemes and conditional probabilities for their pronunciation as specific phonemes are given based on a corpus of 17,310 English words. Phonemes are as given in recent editions of *Webster's New Collegiate Dictionary*, with minor revisions; graphemes are defined as letters or letter clusters corresponding to single phonemes. Grapheme-phoneme probabilities were derived from a revised table of frequency of occurrence of phoneme-to-grapheme correspondences generated in a study of spelling regularities (P. R. Hanna, J. S. Hanna, Hodges, & Rudorf, 1966). This quantitative descriptive information provides an index of the strength of particular grapheme-phoneme associations in English. Suggestions are made for the utilization of these probabilities as estimates of spelling/sound predictability in reading research.

The processes by which readers of a language assign speech sounds to print while reading aloud are of interest to a wide range of researchers. Traditionally part of the domain of educators and other specialists concerned with the acquisition of literacy, the topic has recently gained attention from cognitive psychologists who focus on the processes of lexical access and identification (e.g., Seidenberg, Waters, Barnes, & Tanenhaus, 1984), from neuropsychologists interested in patterns of breakdown in the reading process of formerly literate adults (Coltheart, 1985), and from computer scientists concerned with the development of automatic systems for the pronunciation of print (Allen, 1980).

Researchers have not determined the extent to which reading aloud is based on rule-governed, print-to-sound translation procedures or is accomplished by direct consultation of a mental lexicon in which word pronunciations are listed. In experimental attempts to study this

issue, researchers (e.g., Parkin, 1982) frequently manipulate the orthographic *regularity* of stimulus words, that is, the extent to which their pronunciations can be derived by print-to-sound correspondence rules. Other researchers have manipulated the size of the print segment to be mapped to sound and have focused on the consistency with which sublexical segments are pronounced in real words (Patterson & Morton, 1985; Seidenberg et al., 1984). Although several attempts have been made to characterize the regularities found in print-to-sound mapping (e.g., Venezky, 1970), no comprehensive set of translation rules exists in English. Moreover, objective information is lacking on the extent of association between specific letters (and letter clusters) and their pronunciations in words.

The present study provides a quantitative description of the print-to-sound correspondences that occur in a large corpus of English words. As part of the development of a connectionist model of the oral reading process (see Reggia & Berndt, 1986), probabilities were derived for the correspondences between individual graphemes (i.e., letters or letter clusters corresponding to a single phoneme) and their phonemic realizations. These probabilities are context independent; that is, they do not systematically reflect the word-specific morphological, syllabic, and suprasegmental information that has been argued to constitute an important aspect of the rules that are used to map from print to sound (Henderson, 1985; Venezky, 1970). Thus, the probabilities provide a conservative estimate of the extent to which particular letters and letter clusters are pronounced as particular phonemes in English, but they provide no information about the rules responsible for the derivation of these correspondences.

This project was supported by Grant N00014-85-K-0390 from the Office of Naval Research. R. S. Berndt is supported by Grant KO4-NS-00851 from the National Institute of Neurological and Communicative Disorders and Stroke. The authors are grateful to Patricia Grady, Anne Haendiges, Gloria Kimbles, Michael Morris, and Aita Salasoo for their assistance with various aspects of this project, and also to Ruth D. Bass for careful typing of the manuscript and tables.

The idea of using the corpus generated from P. R. Hanna, J. S. Hanna, Hodges, and Rudorf (1966) for the purposes set out in this study was suggested to us by Max Coltheart.

All three authors are affiliated with the Department of Neurology; James A. Reggia is also with the Department of Computer Science. Reprint requests should be sent to Rita S. Berndt, Department of Neurology, University of Maryland Medical School, 22 S. Greene St., Baltimore, MD 21201.

THE PHONEME-TO-GRAPHEME CORPUS

In 1966, the Office of Education of the United States Department of Health, Education, and Welfare published an analysis of phoneme-to-grapheme correspondences in a corpus of 17,310 words (P. R. Hanna, J. S. Hanna, Hodges, & Rudorf, 1966).¹ The goal of this massive undertaking was to determine the extent to which the spelling of words in English is derivable from their pronunciations. Graphemes were defined as letters or letter clusters corresponding to a single phoneme. Hence there was a mandated one-to-one correspondence between graphemes and phonemes in words. Phonemic realizations for words were obtained from *Webster's New Collegiate Dictionary* (1961, 6th edition) and modified as noted below. A computer program was developed to search the printed spelling of each word to determine the frequency with which each phoneme was represented by a particular graphemic option in the corpus. This procedure generated several large tables of frequencies of occurrence of specific phoneme-to-grapheme correspondences and even larger appendices listing the words in which particular spellings were found. This exercise led to the conclusion that spelling is predictable from pronunciation approximately 80% of the time, under certain conditions.

Although it does not appear that this study was ever put to the educational purposes for which it was intended, it has generated considerable interest among researchers (e.g., Goodman & Caramazza, 1986) studying acquired disorders of spelling. However, its utility for those interested in *reading* has been limited by the fact that the correspondences were generated to predict *spelling*. The probability that a particular grapheme is pronounced as a particular phoneme is not directly retrievable from information concerning the frequency with which a particular phoneme is spelled as a particular grapheme. For example, although the phoneme \ɔ\ is only rarely (probability = .15) spelled AU (as in *auCTION*), the graphemic option AU is almost always ($p = .95$) pronounced \ɔ\. The purpose of this study is to provide probability estimates for the pronunciation of particular graphemes. This was accomplished by generating reverse probabilities for a revised set of the frequencies of occurrence listed by P. R. Hanna et al. (1966; this study is hereinafter referred to as HHH&R).

REVISIONS TO THE HHH&R CORPUS

The grapheme-to-phoneme probabilities were computed on the basis of Tables 17 (vowels) and 18 (consonants) from HHH&R, which list the number and percentage of occurrences of phoneme-to-grapheme associations by the position of the phoneme in syllables (initial, medial, final). For the present study, the position-within-syllable distinction was not used, since information about syllabic structure cannot be presumed to be contained in a word's

graphemic shape prior to the assignment of a pronunciation.

Revisions Related to Pronunciation

Several modifications were made to the phoneme-to-grapheme listings contained in Tables 17 and 18 of HHH&R to bring the pronunciations contained in the corpus in line with contemporary usage. *Webster's New Collegiate Dictionary* (6th edition), which was the basis for the pronunciations generated for the HHH&R corpus, contained several phonemic distinctions, as well as some word pronunciations, that were changed in the 7th (first published in 1963) and subsequent editions. The most important among these changes is the reduction of vowel phonemes from 33 in the 6th edition to 28 in the 7th and later editions. Although HHH&R regrouped the 33 vowels in the 6th edition to a more manageable 22, they retained 4 vowels (all used before the phoneme /r/) that have not been distinguished in recent editions. In addition, the phonemic distinctions carried by these rhotacized vowels are not always apparent, at least to the ear of the contemporary speaker of Eastern American English. For example, the initial vowels in the words *or* and *also* were classified as the same phoneme, whereas the vowels of *chief* and *pier* were distinguished.² In the present study, 17 vowels are used, based on those used in *Webster's New Collegiate Dictionary* (1981, 8th edition).³ Appendix A presents a complete listing of the phonemes used in this study, their equivalents in *Webster's*, in HHH&R, and in the International Phonetic Alphabet with an example word.⁴

As shown in Appendix A, several "phonemes" used by HHH&R (and included in the present study) are actually compounds of two phonemes (e.g., /ks/, /kw/). These are not listed as individual phonemes in the 8th edition of *Webster's*, but are necessary to maintain a one-to-one correspondence between graphemes and phonemes (e.g., X→ks; U→yu).⁵ Such representation by combining phonemes is consistent with the method used in the International Phonetic Alphabet to express single phonemes without dictionary representations (e.g., /ə/ in *puddle*).

A second significant change to the HHH&R corpus that occurred as a result of changes in recent editions of *Webster's* involved the phonemic expression of terminal Y, EY, and IE, as in the words *baby*, *money*, and *movie*. The 6th edition assigns a short-i pronunciation to the final vowel, (e.g., ba'bi), possibly to convey reduced stress. Subsequent editions, and current usage, assign a long-e pronunciation (e.g., ba'bē). Consequently, HHH&R found a high frequency of correspondence between the phoneme 'ih' and the graphemes Y (1,647 occurrences), EY (40), and IE (16) in syllable-final position that would not be reflected in the current dictionary. Adjustment was made by decreasing the ih→Y, IE, EY correspondences by the number of times the former correspondences were found in the spelling corpus (Appendix E, HHH&R), and increasing the figures for cor-

responsibility between those graphemes and the phoneme 'ee' by the same amounts.⁶

Problems with the Letter X

The letter X is typically pronounced as two phonemes: /ks/ or (less frequently) /gz/. This fact was recognized by HHH&R, and /ks/ was included as a phonemic option, whereas /gz/ was not. It appears that words with X→gz were included in the HHH&R corpus separately as X→g and X→z. The same words are included in Appendix E of HHH&R as examples of each correspondence. In addition, the associations X→k and X→s are included in addition to X→ks, and the assignment of these correspondences is based presumably on syllabic considerations. For example, *flex* is segmented as X→ks, while *flaxen* is segmented X→k, X→s. In order to simplify this complex and rather inconsistent set of correspondences, we deleted the associations X→g, X→z, X→k, and X→s and created two larger sets of associations: X→gz and X→ks.⁷

TESTING THE HHH&R CORRESPONDENCES

We tested the HHH&R correspondences against a corpus of words in an effort to determine the degree of difficulty involved in (1) forcing graphemes and phonemes into a one-to-one correspondence, and (2) providing pronunciations for irregularly spelled words. The Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982) was chosen as the test corpus because it contains a relatively large (1,080) set of two-syllable words with a wide range of frequencies, not preselected on the basis of orthographic structure. Words were first converted into keyboard-compatible phonemic equivalents of the pronunciations found for each word in *Webster's New Collegiate Dictionary* (8th edition). If there was more than one dictionary entry, the following rules applied:

1. Only one pronunciation was used unless the alternate pronunciation changed the word's grammatical class (e.g., *abuse*). In the latter condition, both pronunciations were represented separately.
2. For alternate pronunciations that did not change the word class, the first entry was used. In six instances, the secondary dictionary pronunciation was chosen to maintain a one-to-one correspondence between graphemes and phonemes. This was usually necessary when the preferred pronunciation required a silent letter (e.g., *boundary*, *several*) and the secondary pronunciation allowed the sounding of that letter.

Syllable divisions were consistent with the entry in the dictionary, except when the syllable divisions occurred within a phoneme. In these cases, the divided phoneme was recombined before the syllable markers to maintain the one-to-one correspondence between graphemes and phonemes in words containing the grapheme X. This deviation from the dictionary entry was necessary in 11% of entries.

Once the phonemes were determined for each word, the printed word was segmented into graphemes so that a single phoneme corresponded to a single grapheme. With the changes made to HHH&R discussed above, it was possible to express the great majority of words in the Toronto Word Pool without difficulty using the HHH&R correspondences.⁸ This exercise convinced us that although some of the graphemic options used by HHH&R appear arbitrary, the correspondences themselves can account for the pronunciation of a wide sample of words.

Computing the Reverse Probabilities for Print-to-Sound Mapping

In order to apply the extensive correspondences available from the HHH&R study to an investigation of reading, we had to calculate the reverse probabilities with which particular graphemes are realized as particular phonemes (unavailable in HHH&R). For each individual grapheme g, we identified all correspondences between any phoneme p and g. Adding the frequencies of all of these correspondences gave the total frequency of grapheme g. Dividing the frequency of each of these correspondences by the total frequency of g gave an estimate of the desired reverse probabilities, $P(p|g)$. Dividing the total frequency of g by the total number of occurrences of all graphemes gave an estimate of the prior probability of g, $P(g)$. The resulting probabilities are given in Appendix B.

All calculations were derived from a computer file of data based on Tables 17 and 18 of HHH&R. To minimize the possibility of errors, these calculations were done by a computer program written specifically for this purpose. Furthermore, in addition to multiple manual checks, a "parsing" program that used a wealth of redundant information in the data file was used to identify data errors automatically. For example, the parser checked that the total frequency for any phoneme equaled the sum of the frequencies of all correspondences in which that phoneme participated. The parser uncovered several minor arithmetic errors in Tables 17 and 18 of HHH&R.

It should be noted that the overall frequency of occurrence of a particular phoneme, which is used as a basis for calculating these probabilities, represents the number of times a word in the HHH&R corpus contained each phoneme. That is, it does *not* represent the overall phoneme frequency in the language because it is not weighted to reflect the frequency of occurrence of individual words in the language. For example, the phoneme 'th+' has a relatively low overall frequency of occurrence in this study since it occurred in only 149 unique words. Yet because the frequency of occurrence in the language of many of these words (e.g., *the*, *there*) is high, the real frequency of the phoneme 'th+' (and hence of its corresponding grapheme TH) is much higher than estimated in terms of how often particular correspondences are encountered by people reading text. Since some applications of these probabilities (e.g., education, neuropsychology) might want to consider the importance of these different aspects of

frequency of occurrence, we computed the correlation between the frequency of occurrence of each phoneme in the HHH&R corpus and the general frequency of occurrence of phonemes in the language. The second measure was obtained from Denes (1963), who tabulated phoneme frequency from a spoken corpus of 23,052 words from conversational and narrative speech samples. Appendix C contains the rank order of phonemes from our modification of the HHH&R corpus, together with the ranks obtained by Denes. Several of our "combined" phonemes were not included in the Denes corpus and are excluded from the ranking. The Spearman rank-order correlation between the two was .65 ($p < .05$), indicating moderate positive correlation between the frequency of phonemes in individual words and the frequency with which phonemes occur in spoken language. The phonemes that contribute most to lowering this correlation (e.g., 'th+', 'w') can be seen in Appendix B.

DISCUSSION

The probabilities contained in Appendix B provide an estimate of grapheme-phoneme associations as they occur in a corpus of real words. They can be used in a variety of ways to motivate the selection of stimulus words and the construction of pseudowords for use in experiments on reading. One metric based on these associations that could be used to express the overall predictability of a word's pronunciation is the average of the ratio of each probability $P(p|g)$ and the most probable correspondence for that combination. Let P_g^* = the probability of the most frequent correspondence for a grapheme g (e.g., $P_A^* = .54$). Assume that a printed word " $g_1, g_2, \dots, g_n$ " is spoken as phonemes " $p_1, p_2, \dots, p_n$ ". An overall metric of predictability could be computed as

$$m = \frac{1}{n} \sum_{i=1}^n \frac{P(p_i|g_i)}{P_{g_i}^*}.$$

For example, the metric of predictability for the word *tap* would be calculated as follows:

$$T \rightarrow t \quad (P(t|T) = .97, P_T^* = .97)$$

$$A \rightarrow ae \quad (P(ae|A) = .54, P_A^* = .54)$$

$$P \rightarrow p \quad (P(p|P) = 1.00, P_P^* = 1.00)$$

$$m = \frac{1}{3} \left(\frac{.97}{.97} + \frac{.54}{.54} + \frac{1.0}{1.0} \right) = 1.0$$

That is, the pronunciation of the word *tap* is highly predictable from its spelling. The same metric for the word *who* would be calculated:

$$WH \rightarrow h \quad (P(h|WH) = .152, P_{WH}^* = .847)$$

$$O \rightarrow oo \quad (P(oo|O) = .006, P_O^* = .314)$$

$$m = \frac{1}{2} \left(\frac{.152}{.847} + \frac{.006}{.314} \right) = .09$$

The pronunciation of the word *who* is quite unpredictable from its spelling.

For some applications of these data, it might be important to consider the overall probability of occurrence of the graphemes in a word when computing a metric of predictability. For example, the pronunciation of *calf* is highly predictable ($m = 1.0$) using the metric presented above. Yet because the grapheme LF occurs infrequently ($p = .00008$), the correspondence LF $\rightarrow$ f, though a consistent pronunciation for that grapheme, may not be easily learned by beginning readers or well retained by adults with reading disorders. There are a variety of ways in which the predictability metric could be weighted to reflect the probability of occurrence of the graphemes that make up a word.

Much recent research on the question of print-to-sound mapping has divided words into sets that are regular/irregular and/or consistent/inconsistent. The latter dichotomy is intended to express the actual occurrence of print-to-sound associations in words, independent of whether or not a word is said to have a regular spelling (i.e., derivable by rule). For example, the word *seen* is said to be regular but inconsistent, since there are other words with the same three-letter ending (e.g., *been*) that are pronounced differently (Glushko, 1979). In contrast, *seed* is said to be regular and consistent, since all words ending in "eed" are pronounced like *seed*. It is argued that inconsistent words require more processing time than do consistent words because of interfering activation from similarly spelled words with different pronunciations. This characterization of pronunciation consistency rests on a concept of word segments that is somewhat unclear. Would the sharing of the initial three letters rather than the final three cause comparable interference (see Patterson & Morton, 1985)?

The probabilities contained in Appendix B provide an alternate way of looking at this question, while avoiding the problem of specifying these supraphemic but sublexical segments. Information about inconsistency of grapheme-phoneme correspondences obtained from Appendix B can be used to supplement these divisions of words into consistent and inconsistent groups. For example, Seidenberg et al. (1984) constructed pairs of one-syllable words with the same three-letter endings, one of which has an exceptional pronunciation, and the other of which is regular but inconsistent by virtue of the existence of the exceptional member of the pair. Focusing on the vowels of these words rather than on the final three letters, one can see that some of the inconsistent words are more inconsistent than others. Independent of the terminal letter of the word pair, the vowel of the inconsistent word *braid*, for example, is likely to be pronounced

'ay' ($p = .734$) and that of its matched exception word *plaid* is quite unlikely to be pronounced 'ae' ($p = .003$). Other pairs show different patterns. The pronunciation of the vowel of the regular inconsistent word *pool* ($p = .570$) is only marginally more probable than is the pronunciation of its matched exception *wool* ($p = .376$). It is not yet known to what extent these considerations might prove to be important in subjects' performance of reading-related tasks, but the availability of these probabilities provides a means for evaluating empirically the role of grapheme-phoneme consistency in studies of word segment consistency.

The usefulness of these probabilities will be limited to cases for which context-free estimates of association are required. Many difficult and important context-dependent questions about how reading is accomplished have been specifically avoided here, the most important of which is the actual segmentation of printed words into units to which a pronunciation can be assigned. Clearly, other divisions of printed words than those employed here would have resulted in a different set of correspondences. In fact, many alternate possibilities are suggested by the probabilities generated in Appendix B. For example, graphemic options with low prior probability (e.g., CHT→t, $p = .00001$) occur in very few words and should probably not be thought of as meaningful sublexical segments at all, but as word-specific orthographic anomalies. Many of the graphemes involving a silent member might be productively thought of in this way. In fact, the entire treatment of silent letters by HHH&R, as reflected in Appendix B, might be handled differently if the goal is to describe segments that readers actually use. Graphemes combining vowels with silent E (e.g., A-E, U-E, I-E), originally intended to be limited to cases in which the two vowels were separated by only one consonant grapheme (HHH&R, p. 33), were ultimately used in words in which more than one consonant grapheme separated the two vowels (e.g., *orange*, *advance*). This usage prevented any systematic assumption of a vowel-lengthening function for final silent E, and also obviated the need for other combination graphemes involving silent E (e.g., GE and CE in the last examples). These latter possibilities (not included among the HHH&R options) would seem to be reasonable candidates for the type of rule-governed segments that readers might actually use to assign pronunciations to print.

The set of probabilities of association provided in this paper may serve as a point of departure for the development of these and other aspects of alternative systems for investigating the translation of print to sound.

REFERENCES

- ALLEN, J. (1980). Speech synthesis from text. In J. C. Simon (Ed.), *Spoken language generation and understanding* (pp. 383-396). Dordrecht, The Netherlands: Reidel.
- COLTHEART, M. (1985). Cognitive neuropsychology and the study of reading. In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI* (pp. 3-37). Hillsdale, NJ: Erlbaum.
- DENES, P. B. (1963). On the statistics of spoken English. *The Journal of the Acoustical Society of America*, 35, 892-904.
- FRIENDLY, M., FRANKLIN, P. E., HOFFMAN, D., & RUBIN, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, 14, 375-399.
- GLUSHKO, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception & Performance*, 6, 674-691.
- GOODMAN, R. A., & CARAMAZZA, A. (1986). Phonologically plausible errors: Implications for a model of the phoneme-grapheme conversion mechanism in the spelling process. In G. Augst (Ed.), *Proceedings of the International Colloquium on Graphics and Orthography* (pp. 300-325). Siegen, W. Germany.
- HANNA, P. R., HANNA, J. S., HODGES, R. E., & RUDOLF, E. H. (1966). *Phoneme-grapheme correspondences as cues to spelling improvement*. Washington, DC: U.S. Department of Health, Education, and Welfare.
- HENDERSON, L. (1985). Issues in the modelling of pronunciation assembly in normal reading. In K. E. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading*. London: Erlbaum.
- PARKIN, A. J. (1982). Phonological recoding in lexical decision: Effects of spelling-to-sound regularity depend on how regularity is defined. *Memory & Cognition*, 10, 43-53.
- PATTERSON, K. E., & MORTON, J. (1985). From orthography to phonology: An attempt at an old interpretation. In K. E. Patterson, J. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading*. London: Erlbaum.
- REGGIA, J. A., & BERNDT, R. S. (1986). Modelling reading aloud and its relevance to acquired dyslexia. *Computer Methods & Programs in Biomedicine*, 22, 13-19.
- SEIDENBERG, M. S., WATERS, G. S., BARNES, M. A., & TANENHAUS, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning & Verbal Behavior*, 23, 383-404.
- VENEZKY, R. L. (1970). *The structure of English orthography*. The Hague: Mouton.
- Webster's New Collegiate Dictionary* (6th ed.). (1961). Springfield, MA: Merriam-Webster.
- Webster's New Collegiate Dictionary* (8th ed.). (1981). Springfield, MA: Merriam-Webster.

NOTES

1. Reprints of this study are available at reproduction cost from ERIC (Educational Resources Information Center), Washington, DC 20208.
2. It should be noted that the problem of r-influenced vowels has not been solved by our modification of the HHH&R phonemes. Although recent editions of *Webster's* do not have a separate symbol for vowels preceding /r/, many dictionary entries contain (/ə/) between the vowel and /r/ to express the influence of /r/. We considered this parenthetical schwa to be a secondary pronunciation option; thus we excluded it to maintain consistency of interpretation across phonemes with parenthetical elements. Nonetheless, some of the pronunciations produced by this set of limited possibilities appeared to distort the influence of the /r/ (e.g., *force* → fōrs).
3. We excluded five non-English phonemes (e.g., French [œ] and German [uē]), two stress distinctions ('ər; 'ē), a phoneme distinction not included in *Webster's* 6th edition but added in later editions (yū as in *fury*), and a dialectal variant not realized in Eastern American English (/ā/ for *father* to differentiate it from *bother*).
4. Throughout this paper, phonemes are expressed in lowercase in the keyboard-compatible phonemic symbol of Appendix A (when embedded in text, these are shown in single quotation marks). Graphemes are expressed in uppercase letters. Whole words are italicized. Symbols from the International Phonetic Alphabet are bounded by slashes.
5. The maintenance of a one-to-one correspondence between graphemes and phonemes is required by the definition of *grapheme* used in the HHH&R study: graphemes are defined as segments of characters that correspond to single phonemes. Other concepts of the theoretic-

cal entity *grapheme* do not imply such correspondence (see Henderson, 1985, for discussion). The question of how readers of a language segment a character string into units that map onto phonemes is an important empirical question that is not addressed in this paper.

6. Two additional changes were made for the same reason, each of which occurred in only one word of the corpus. The correspondence 'ih' → AY (*yesterday*) was changed to 'ay' → AY; 'ih' → EA (*guinea*) was changed to 'ee' → EA to conform to *Webster's* (8th edition) pronunciations of those words. One additional grapheme (LN) was deleted from the corpus altogether, as it had apparently been included erroneously (i.e., LN → /l/, as in *kiln*).

7. Although this solution appeared to account well for all of the words in HHH&R's Appendix E that contained X, there are other words that may require $X \rightarrow k$ (e.g., *except*). An additional set of divided correspondences with X ($X \rightarrow k$; $X \rightarrow sh$, as in *luxury*) was also deleted. Since only three words from the corpus contained this correspondence, it was not judged appropriate to create a new phonemic category /ksh/ to cover this occurrence. It could also be argued that /gz/ should be added to account for *luxurious* (see Webster's, 6th edition), but we did not do so. Similar problems are apparent with mappings to the combined phoneme /kw/, but were not as pronounced since the typical graphemic option involved two characters (QU). Nonetheless, the rationale for the correspondences chosen by HHH&R to express graphemes Q, OU, and COU is not always clear.

8. Words in the Toronto Word Pool that could not be pronounced using correspondences in HHH&R were *acre*, *business*, *colonel*, *disguise*, *goodbye*, *handsome*, and *iron*.

Appendix A

Phoneme Equivalents

<i>Webster's New Collegiate Dictionary</i>	Keyboard Compatible Phonemic Symbol	HHH&R Symbol	International Phonetic Alphabet	Example
ā	ay	A	/e/	<u>a</u> le
a	ae	A3	/æ/	<u>a</u> dd
ē	ee	E, E2	/i/	<u>bee</u>
e	eh	E3, A2	/ɛ/	<u>e</u> nd
ēr	er	E5, U2	/ɛr/	<u>fa</u> ther
ī	ai	I	/ai/	<u>hi</u> gh
i	ih	I3	/I/	<u>bi</u> n
ō	o	O	/o/	<u>bo</u> at
ā	ah	O3, A5	/a/	<u>c</u> ot
ò	aw	O5, O2	/ɔ/	<u>s</u> oft
ü	oo	O6	/u/	<u>fo</u> od
ù	u	O7	/u/	<u>h</u> ook
yü	yu	U	/ju/	<u>un</u> ite
`ə	uh+	U3	/ʌ/	<u>u</u> p
oi	oy	OI	/oi/	<u>bo</u> y
au	au	OU	/əʊ/	<u>ou</u> t
ə	uh—	ə	/ə/	<u>ab</u> out
b	b	B	/b/	<u>b</u> ut
d	d	D	/d/	<u>d</u> ay
f	f	F	/f/	<u>f</u> an
g	g	G	/g/	<u>g</u> o
h	h	H	/h/	<u>h</u> at
j	dj	J	/dʒ/	<u>j</u> oke
k	k	K	/k/	<u>k</u> ee <u>p</u>
l	l	L	/l/	<u>l</u> ate
m	m	M	/m/	<u>m</u> an
n	n	N	/n/	<u>n</u> od
p	p	P	/p/	<u>p</u> en
r	r	R	/r/	<u>r</u> at
s	s	S	/s/	<u>s</u> ue
t	t	T	/t/	<u>t</u> wo
v	v	V	/v/	<u>v</u> an
w	w	W, HW	/w/	<u>w</u> ait,

				what
y	y	Y	/j/	<u>yet</u>
z	z	Z	/z/	<u>zone</u>
ch	tch	CH	/tʃ/	<u>chin</u>
--	ks	KS	/ks/	<u>ox</u>
--	gz	--	/gz/	<u>exist</u>
--	kw	KW	/kw/	<u>quit</u>
ʔl	ul	Ll	/l/	<u>puddle</u>
ʔm	um	Ml	/m/	<u>chasm</u>
ʔn	un	Nl	/n/	<u>pardon</u>
ŋ	ng	NG	/ŋ/	<u>sing</u>
sh	sh	SH	/ʃ/	<u>she</u>
th	th—	T1	/θ/	<u>thin</u>
th	th+	T2	/ð/	<u>then</u>
zh	zh	ZH	/ʒ/	<u>rouge</u>
--	nl	H9	--	<u>honest</u>

Note—HHH&R symbols are from *Phoneme-Grapheme Correspondences as Cues to Spelling Improvement* by P. R. Hanna, J. S. Hanna, R. E. Hodges, and E. H. Rudorf, 1966, Washington, DC: U.S. Department of Health, Education, and Welfare.

Appendix B

Conditional Probabilities for Grapheme-to-Phoneme Correspondences, with Example Words*

Correspondences with Example Words				
Grapheme†	Prior Probability	Phoneme	Conditional Probability	Example
Vowels				
A	.0712	ae	.542	<u>cab</u>
		uh—	.186	<u>canal</u>
		ay	.129	<u>angel</u>
		ah	.077	<u>wad</u>
		aw	.021	<u>also</u>
		er	.021	<u>coward</u>
		eh	.020	<u>many</u>
		ih	.0005	<u>spinach</u>
A-E	.0111	ay	.651	<u>ate</u>
		ih	.154	<u>senate</u>
		ae	.121	<u>madame</u>
		eh	.042	<u>orange</u>
		ah	.025	<u>are</u>
		aw	.002	<u>false</u>
		uh—	.002	<u>pirate</u>
AE	.00005	ee	.833	<u>algae</u>
		eh	.166	<u>aesthetic</u>
AH	.00003	ah	1.000	<u>dahlia</u>
AI	.0026	ay	.734	<u>aid</u>
		eh	.176	<u>said</u>
		ih	.053	<u>captain</u>
		uh—	.031	<u>villain</u>
		ae	.003	<u>plaid</u>
AI-E	.0002	ay	.818	<u>aide</u>
		eh	.136	<u>millionaire</u>
		ai	.045	<u>aisle</u>
AIGH	.00003	ay	1.000	<u>straight</u>
AO	.00001	aw	1.000	<u>extraordinary</u>
AU	.0014	aw	.948	<u>faucet</u>
		ah	.025	<u>aunt</u>
		o	.019	<u>taupe</u>
		uh—	.006	<u>epaulet</u>
AU-E	.0001	aw	.750	<u>cause</u>
		o	.166	<u>mauve</u>
		ay	.083	<u>gauge</u>
AUGH	.0001	aw	1.000	<u>caught</u>

APPENDIX B (Continued)									
AW	.0006	aw	1.000	<u>law</u>	EY-E	.00006	ai	1.000	<u>eye</u>
AW-E	.00001	aw	1.000	<u>awe</u>	I	.0688	ih	.716	<u>in</u>
AY	.0012	ay	.970	<u>way</u>			uh-	.180	<u>civil</u>
		ai	.022	<u>kayak</u>			ai	.074	<u>find</u>
		eh	.007	<u>says</u>			er	.015	<u>affirm</u>
AY-E	.000009	ay	1.000	<u>aye</u>	I-E	.0086	y	.008	<u>senior</u>
E	.0730	eh	.419	<u>end</u>			ee	.005	<u>ski</u>
		er	.249	<u>baker</u>			ai	.589	<u>ice</u>
		ee	.230	<u>senior</u>			ih	.356	<u>give</u>
		uh-	.096	<u>rebel</u>			ee	.046	<u>elite</u>
		ay	.002	<u>cafe</u>	IA	.00001	uh-	.006	<u>fragile</u>
		ah	.0006	<u>encore</u>	IA-E	.00002	er	.001	<u>dirge</u>
		ih	.0006	<u>pretty</u>	IE	.0011	uh-	1.000	<u>parliament</u>
		y	.0001	<u>azalea</u>			ih	1.000	<u>marriage</u>
E-E	.0032	eh	.321	<u>edge</u>			ee	.492	<u>chief</u>
		uh-	.286	<u>science</u>			ai	.203	<u>tried</u>
		ee	.252	<u>eve</u>			uh-	.171	<u>ancient</u>
		er	.119	<u>traverse</u>	IE-E	.0002	ih	.101	<u>mischief</u>
		ay	.017	<u>fete</u>			eh	.031	<u>friend</u>
		ih	.002	<u>privilege</u>			ee	.838	<u>piece</u>
EA	.0047	ee	.576	<u>eat</u>			uh-	.129	<u>patience</u>
		eh	.298	<u>head</u>	IEU	.00003	ih	.032	<u>sieve</u>
		er	.056	<u>earn</u>	IEW	.00005	yu	1.000	<u>lieu</u>
		ah	.035	<u>heart</u>	IGH	.0008	yu	1.000	<u>view</u>
		ay	.027	<u>break</u>	O	.0550	ai	1.000	<u>night</u>
		uh-	.005	<u>pageant</u>			o	.314	<u>only</u>
EA-E	.0003	ee	.882	<u>ease</u>			uh-	.269	<u>carton</u>
		er	.058	<u>hearse</u>			ah	.261	<u>odd</u>
		eh	.029	<u>mileage</u>			aw	.072	<u>off</u>
		ih	.029	<u>cleanse</u>			er	.053	<u>honor</u>
EAU	.0001	o	.545	<u>beau</u>			uh+	.018	<u>oven</u>
		yu	.454	<u>beauty</u>			oo	.006	<u>who</u>
EE	.0026	ee	.979	<u>bee</u>	OO-E	.0001	u	.002	<u>today</u>
		ih	.020	<u>been</u>	O-E	.0043	ih	.0001	<u>women</u>
EE-E	.00008	ee	1.000	<u>geese</u>			oo	1.000	<u>choose</u>
EI	.0005	ee	.315	<u>ceiling</u>			o	.785	<u>code</u>
		ay	.245	<u>vein</u>			uh+	.055	<u>come</u>
		ih	.192	<u>forfeit</u>			uh-	.044	<u>welcome</u>
		ai	.105	<u>stein</u>			aw	.044	<u>gone</u>
		eh	.105	<u>heifer</u>			ah	.042	<u>dodge</u>
		uh-	.035	<u>sovereign</u>			oo	.025	<u>lose</u>
EI-E	.00007	ee	.750	<u>seize</u>	OA	.0012	er	.002	<u>worse</u>
		ay	.250	<u>seine</u>			o	.933	<u>coat</u>
EIGH	.0001	ay	.857	<u>eight</u>	OA-E	.00002	aw	.066	<u>broad</u>
		ai	.142	<u>height</u>	OE	.0002	o	1.000	<u>coarse</u>
EO	.0001	uh-	.666	<u>pigeon</u>			ee	.590	<u>doe</u>
		eh	.200	<u>leopard</u>			oo	.227	<u>amoeba</u>
		ee	.133	<u>people</u>	OH	.00003	oo	.181	<u>canoe</u>
EOU	.00007	uh-	1.000	<u>courageous</u>	OI	.0008	o	1.000	<u>ohm</u>
ET	.00008	ay	1.000	<u>beret</u>	OI-E	.00009	oy	1.000	<u>coil</u>
EU	.0003	yu	.717	<u>feud</u>			oy	.800	<u>voice</u>
		er	.153	<u>amateur</u>			uh-	.200	<u>tortoise</u>
		oo	.102	<u>sleuth</u>	OO	.0027	oo	.570	<u>boot</u>
		u	.025	<u>pleurisy</u>			u	.376	<u>wood</u>
EU-E	.000009	yu	1.000	<u>deuce</u>			o	.029	<u>door</u>
EW	.0005	yu	.603	<u>news</u>	OU	.0064	uh+	.023	<u>blood</u>
		oo	.349	<u>shrewd</u>			uh-	.480	<u>enormous</u>
		o	.047	<u>sew</u>			au	.324	<u>out</u>
EW-E	.000009	yu	1.000	<u>ewe</u>			uh+	.042	<u>touch</u>
EY	.0005	ee	.741	<u>key</u>			o	.041	<u>four</u>
		ay	.225	<u>they</u>			oo	.041	<u>you</u>
		ai	.016	<u>geyser</u>			u	.035	<u>could</u>
		eh	.016	<u>eyrie</u>	OU-E	.0006	er	.031	<u>glamour</u>
							w	.001	<u>bivouac</u>
							au	.794	<u>ounce</u>

APPENDIX B (Continued)

APPENDIX B (Continued)											
OUGH	.0002	o	.147	<u>course</u>	CHT	.00001	t	1.000	yacht		
		oo	.044	<u>route</u>	CI	.0007	sh	1.000	social		
		er	.014	<u>scourge</u>	CK	.0026	k	1.000	cracked		
		aw	.517	<u>bought</u>	CQ	.00002	k	1.000	acquiesce		
		o	.275	<u>dough</u>	CQU	.00004	kw	1.000	acquaint		
		au	.137	<u>drought</u>	CS	.0002	ks	1.000	ethics		
		oo	.068	<u>through</u>	CT	.00001	t	1.000	indict		
OW	.0022	o	.502	<u>own</u>	CZ	.000009	z	1.000	czar		
		aw	.481	<u>towel</u>	D	.0336	d	.991	dad		
		ah	.016	<u>knowledge</u>			dj	.008	educate		
OW-E	.00002	au	.666	<u>browse</u>	DD	.0006	d	1.000	add		
		o	.333	<u>owe</u>	DG	.0004	dj	1.000	budget		
OY	.0004	oy	.972	<u>boy</u>	DI	.00001	dj	1.000	cordial		
		ai	.020	<u>coyote</u>	DJ	.0001	dj	1.000	adjust		
OY-E	.000009	oy	1.000	<u>gargoyle</u>	ED	.0002	t	1.000	paced		
		uh+	.417	<u>up</u>	EL	.0001	ul	1.000	easel		
U	.0267	yu	.280	<u>union</u>	EN	.0007	un	1.000	shorten		
		uh-	.102	<u>medium</u>	ES	.0004	z	1.000	goes		
		er	.080	<u>survey</u>	F	.0146	f	.998	farm		
		u	.068	<u>put</u>			v	.001	of		
		oo	.032	<u>truth</u>	FF	.0016	f	1.000	cuff		
		w	.016	<u>liquid</u>	FT	.00002	f	1.000	often		
		eh	.0006	<u>bury</u>	G	.0169	g	.640	gap		
		ih	.001	<u>busy</u>			dj	.351	gem		
		yu	.703	<u>use</u>			zh	.008	regime		
		oo	.093	<u>rude</u>	GG	.0006	g	.971	egg		
U-E	.0033	er	.090	<u>measure</u>			dj	.028	exaggerate		
		uh+	.063	<u>budge</u>	GH	.0001	g	.555	ghost		
		u	.030	<u>sure</u>			f	.444	rough		
		uh-	.010	<u>ferrule</u>	GI	.0001	dj	1.000	region		
		ih	.008	<u>minute</u>	GM	.00007	m	1.000	phlegm		
		yu	.627	<u>cue</u>	GN	.0002	n	1.000	gnat		
		oo	.372	<u>blue</u>	GUE	.0001	g	1.000	rogue		
		ih	.533	<u>build</u>	H	.0070	h	1.000	hot		
		yu	.266	<u>suit</u>	IL	.00006	ul	1.000	evil		
		oo	.200	<u>fruit</u>	IN	.0003	un	1.000	basin		
UI	.0002	oo	1.000	<u>juice</u>	J	.0020	dj	1.000	jam		
		oo	1.000	<u>buoyant</u>	K	.0055	k	1.000	keep		
UI-E	.00003	oo	1.000	<u>buoyant</u>	KH	.00002	k	1.000	khaki		
UO	.00001	oo	1.000	<u>buoyant</u>	KN	.0003	n	1.000	knee		
UY	.00002	ai	1.000	<u>buy</u>	L	.0451	l	1.000	lad		
Y	.0193	ee	.786	<u>story</u>	LD	.00005	d	1.000	could		
		ai	.100	<u>typist</u>	LE	.0057	ul	1.000	able		
		ih	.073	<u>myth</u>	LF	.00008	f	1.000	calf		
		y	.025	<u>yes</u>	LK	.0001	k	1.000	talk		
		uh-	.010	<u>ethyl</u>	LL	.0045	l	1.000	all		
		er	.002	<u>satyr</u>	LM	.0001	m	1.000	calm		
		ai	.958	<u>type</u>	LV	.00003	v	1.000	calves		
		ih	.041	<u>apocalypse</u>	M	.0313	m	.971	man		
		Consonants							um	.028	chasm
		AL	.00003	ul	1.000	<u>medal</u>	MB	.0002	m	1.000	lamb
B	.0206	b	1.000	<u>bad</u>	MM	.0012	m	1.000	mummy		
BB	.0005	b	1.000	<u>ebb</u>	MN	.00007	m	.875	hymn		
BT	.0001	t	1.000	<u>debt</u>			n	.125	mnemonics		
C	.0420	k	.757	<u>cab</u>	N	.0710	n	.967	nap		
		s	.234	<u>cent</u>			ng	.032	bank		
		sh	.008	<u>ocean</u>	NG	.0033	ng	1.000	length		
		tch	.0004	<u>cello</u>	NGUE	.00001	ng	1.000	tongue		
CC	.0007	k	1.000	<u>occupy</u>	NN	.0011	n	1.000	inn		
CCH	.000009	k	1.000	<u>saccharin</u>	OL	.000009	ul	1.000	pistol		
CE	.00001	sh	1.000	<u>ocean</u>	ON	.00002	un	1.000	pardon		
CH	.0045	tch	.640	<u>chart</u>	P	.0304	p	1.000	pad		
		k	.290	<u>chorus</u>	PB	.000009	b	1.000	raspberry		
		sh	.069	<u>chef</u>	PH	.0022	f	1.000	graph		
					PN	.00002	n	1.000	pneumonia		

APPENDIX B (Continued)

PP	.0014	p	1.000	apple
PS	.0001	s	1.000	psalm
PT	.00001	t	1.000	receipt
Q	.0001	k	1.000	liquid
QU	.0020	kw	.876	quilt
		k	.123	bouquet
R	.0841	r	1.000	rap
RH	.0001	r	1.000	rhyme
RR	.0019	r	1.000	purr
S	.0488	s	.868	sit
		z	.120	design
		zh	.006	usual
		sh	.003	sugar
SC	.0008	s	.865	scene
		sh	.067	crescendo
		z	.033	discern
		k	.033	viscount
SCH	.00003	s	.500	schism
		sh	.500	schwa
SCI	.00004	sh	1.000	conscious
SH	.0036	sh	1.000	ship
SI	.0008	zh	.568	fusion
		sh	.431	pension
SL	.00004	l	1.000	island
SS	.0042	s	.952	boss
		z	.028	dessert
		sh	.019	pressure
SSI	.0004	sh	1.000	mission
ST	.0003	s	1.000	listen
SW	.00003	s	1.000	sword
T	.0713	t	.973	tab
		tch	.022	culture
		sh	.003	initiate
TCH	.0005	tch	1.000	watch
TG	.000009	g	1.000	mortgage
TH	.0051	th-	.732	thaw
		th+	.265	the
		t	.001	thyme
TI	.0076	sh	.983	action
		tch	.015	question
		zh	.001	equation
TT	.0019	t	1.000	putt
TW	.00002	t	1.000	two
V	.0136	v	1.000	yat
W	.0053	w	1.000	we
WH	.0009	w	.847	what
		h	.152	who
WR	.0004	r	1.000	wrap
X	.0033	ks	.885	ox
		gz	.114	exist
Z	.0021	z	.996	zoo
		s	.025	waltz
		zh	.008	azure
ZZ	.0002	z	1.000	buzz
\$H	.0003	null	1.000	heir

*Derived from revised tables of frequency of occurrence of grapheme-to-phoneme correspondences in a corpus of 17,310 words (*Phoneme-Grapheme Correspondences as Cues to Spelling Improvement* by P. R. Hanna, J. S. Hanna, R. E. Hodges, and E. H. Rudorf, 1966, Washington, DC: U.S. Department of Health, Education, and Welfare). †The total number of graphemes in the word corpus was 108,419.

Appendix C

Rank Orders of Phoneme Occurrences in Word Corpus and in Spoken Language

Phoneme	Rank of Occurrence in		Rank of Occurrence in Spoken	
	17,320-Word Corpus*		23,052-Word Narration†	
r	1		12	
t	2		2	
n	3		4	
s	4		5	
ih	5		3	
uh-	6		1	
l	7		7	
k	8		10	
ee	9		30	
ae	10		24	
eh	11		22	
d	12		6	
m	13		8	
p	14		17	
er	15		34	
o	16		18	
b	17		15	
ay	18		25	
ah	19		29	
f	20		19	
sh	21		33	
v	22		16	
ai	23		11	
uh+	24		21	
g	25		28	
z	26		14	
dj	27		36	
aw	28		39	
h	29		20	
w	30		13	
ng	31		27	
tch	32		37	
oo	33		26	
th-	34		35	
au	35		31	
u	36		32	
oy	37		38	
th+	38		9	
y	39		23	
zh	40		40	

*Phoneme-Grapheme Correspondences as Cues to Spelling Improvement by P. R. Hanna, J. S. Hanna, R. E. Hodges, and E. H. Rudorf, 1966, Washington, DC: U.S. Department of Health, Education, and Welfare. †Data from "On the Statistics of Spoken English" by P. B. Denes, 1963, *The Journal of the Acoustical Society of America*, 35, 892-904.