Grapheme–phoneme probabilities in British English

P. F. D. GONTIJO

University of Edinburgh, Edinburgh, Scotland

I. GONTIJO

Heriot-Watt University, Edinburgh, Scotland

and

R. SHILLCOCK

University of Edinburgh, Edinburgh, Scotland

A computational analysis of a large British English database was performed and frequencies of occurrence of grapheme–phoneme correspondences were obtained. A computer program was implemented, which used these frequencies to predict the probabilities of all possible pronunciations of any given string of graphemes. These results led to a proposal for a quantitative method of measurement of the orthographic depth of different languages.

One of the prime goals of cognitive scientists has been the understanding of the mechanisms underlying reading, at the heart of which is the relationship between graphemes and phonemes. In alphabetic scripts, such as English, the basic unit represented by a grapheme is essentially a phoneme (Venezky, 1970). A grapheme is defined as a letter or combination of letters that represents a phoneme. For example, the word *thought* contains seven letters and three graphemes (*th*, *ough*, *t*), representing three phonemes. The relationship between the orthographic and phonological forms of words varies across languages. Some languages, such as Serbo-Croatian (Katz & Feldman, 1981), have a shallow orthography, meaning that the phonological forms of words are transparently represented in the orthography. In shallow orthographies, the grapheme–phoneme relationship tends to be invariant (or consistent), and the pronunciation of a novel word may be accurately predicted by a relatively small set of rules.

In English, however, the correspondence between spelling and sound is not transparent. A phoneme can be realized by many different graphemes (e.g., /a/ in *mine*, *pie*, and *my*), and a grapheme can also be realized by many different phonemes (e.g., the letter *a* in *fate*, *pat*, and *wash*). According to some estimates, 20% of all word types in English violate regularization rules (J. S. Hanna & P. R.

Hanna, 1959). Henderson (1982) points out that from the 3,000 most common words in English, as many as 21% break the rules proposed in Wijk (1966). Thus, English has a “deep orthography.”

This deep orthographic nature has generated a demand for corpora information on the structure of the language. Accordingly, many databases have been created to address particular aspects of the language (Berndt, D’Aurechy, & Reggia, 1994; Berndt, Reggia, & Mitchum, 1987; P. R. Hanna, J. S. Hanna, Hodges, & Rudorf, 1966; Treiman, Bijeljac-Babic, & Richmond-Welty, 1995; Venezky, 1970; Wijk, 1966; Ziegler, Stone, & Jacobs, 1997). The present study deals with grapheme–phoneme correspondences in English and is based on Berndt et al.’s (1987) work on pronunciation probability. They used a corpus of 17,310 words (P. R. Hanna et al., 1966) to provide probability estimates for the pronunciation of particular graphemes in American English. Probabilities were derived for the correspondences between individual graphemes and their phonemic realizations. In the present work, a much larger British English corpus was used, containing 17.9 million word tokens corresponding to 160,595 word types.

In the Berndt et al. (1987) work and in the present research, the calculated probabilities were context independent; that is, they did not systematically reflect word-specific morphological, syllabic, or suprasegmental information. Thus, the probabilities produced were conservative estimates of the extent to which particular letters and letter clusters are pronounced as particular phonemes in English. Also, such probabilities do not provide any information about the rules responsible for the correspondences. Nevertheless, the significance of this type of information resides in its utility in building and testing models of word perception and spelling and speech pro-

I.G. is now at the Department of Electrical Engineering, University of California, Los Angeles. Correspondence should be addressed to P. F. D. Gontijo, who is now at the Department of Psychology, Franz Hall A407, University of California, Los Angeles, CA 90095 (e-mail: pgontijo@ucla.edu).

Note: The authors have a direct financial interest in some of the software described in this paper.—*Editor*

Table 1
Correspondence Between the CELEX (Disk)
and IPA Character Sets

Disk	IPA	Disk	IPA	Disk	IPA	Disk	IPA	Disk	IPA
p	p	r	r	x	x	I	i	u	u:
b	b	f	f	h	h	E	e	3	ɜ:
t	t	v	v	w	w	{	æ	1	eɪ
d	d	T	θ	J	tʃ	V	ʌ	2	aɪ
k	k	D	ð	—	ɔ	Q	ɒ	4	ɔɪ
g	g	s	s	—	ɪ	U	ʊ	5	əʊ
N	ŋ	z	z	F	ɱ	@	ə	6	aʊ
m	m	S	ʃ	H	ɲ	i	i:	7	ɪə
n	n	Z	ʒ	P	ɭ	#	ɑ:	8	ɛə
l	l	j	j	R	*	\$	ɔ:	9	ʊə
		c	æ	q	ã:	~	õ:	0	æ:

duction systems and in the creation of better controlled experimental stimuli (Ziegler et al., 1997).

The Statistical Calculations

All the word frequency information used in this study was taken from the CELEX database (Version 2.5) (Baayen, Piepenbrock, & Gulikers, 1995), representing approximately 17.9 million words. The CELEX database is in table format, with each word and its associated information appearing in a single row, divided into columns. The first column contains a number uniquely identifying the word form, which is given in the second column. The third column gives the frequency of occurrence in the COBUILD corpus, and a later column gives the phonemic transcription of the word form. The database therefore contains the phonemic transcription corresponding to each grapheme present in the English language.

We used a basic search algorithm to take one grapheme at a time and to determine whether it was present in each word form (each row) of the database. When a grapheme was found in a word form, its corresponding phoneme was searched in the phonemic transcription. The frequency of the grapheme, pronounced as the particular phoneme, was then retrieved from the third column. The algorithm will be described more fully below. It is important to emphasize the size of the corpus from which the CELEX database was derived; the previous attempt at deriving grapheme–phoneme correspondences used a corpus of only 17,000 words. The present study produced better results for the less frequent graphemes and contained many additional grapheme–phoneme correspondences. Note also that, in the previous study, the grapheme–phoneme probabilities were obtained from previously computed phoneme–grapheme associations and starting from spoken words. Finally, the phonemic transcriptions in the CELEX database are based on British English, as opposed to American English (which was used in the previous study).

Preliminary Statistics

CELEX has four different sets of characters for phonemic transcription. In this study, the CELEX (“disk”) rep-

resentation of phonemes will be used in what follows. The international phonetic alphabet (IPA) transcription of phonemes will be given in parentheses. In the disk character set, each phoneme is represented by a single ASCII character, with very few exceptions. The most elaborate are the phonemes associated with the grapheme *x*, which is transcribed as *ks*, *gz*, and *ks* (/ks/, /gz/ and /kʃ/), and one of the possible pronunciations of the grapheme *u* when transcribed as *ju* (/ju:/). Table 1 shows the correspondence between the CELEX (disk) characters and the IPA character set.

The total number of lines (number of word forms) in the database is 160,595. The mean length (number of characters) and standard deviation of word types and of their phonemic transcriptions are 8.3 ± 2.7 characters and 6.9 ± 2.4 characters, respectively. Thus, on average, the grapheme representation of a word is longer than its phoneme representation, which is expected in English. It can be seen in Table 2 that the vast majority of word types has two, three, one, or four syllables, in this order. There is only one “word” with 12 syllables: *European economic community*. CELEX lists not only single words but also some expressions, phrasal verbs, abbreviations, hyphenated words, and so on. There are 2,887 lines in the database that did not contain single, isolated words. This figure represents only 1.8% of the whole database. The total number of unique graphemes found was 195, and the total

Table 2
Distribution of Words With
Different Numbers of Syllables in CELEX

Syllables	Words
1	24,913
2	61,737
3	45,652
4	20,210
5	6,245
6	1,433
7	314
8	58
9	22
10	7
11	3
12	1

number of grapheme–phoneme correspondences was 461.

Grapheme Search Strategies

Graphemes composed of three or more letters were searched first, followed by two-letter graphemes, and, finally, single-letter graphemes. This order is important, since it simplified some of the search strategies. The grapheme–phoneme associations from Berndt et al. (1987) were used as a starting point to build our own database of associations. For some of the Berndt et al. (1987) graphemes, new phoneme associations were found and included. Also, various additional graphemes and their phonemic transcriptions were detected. These additions obviously result from the fact that our word corpus was much larger and also from our use of British English.

Searching bigrams, trigrams, and longer graphemes.

Long graphemes, composed of two or more letters, are easier to search, since they generally occur only once in a word. Therefore, prior to counting the frequency of occurrence of a given long grapheme, a simple program was run to check whether it appeared more than once in any of the words of the database. For example, only three trigrams were found to repeat in words: *igh*, *sch*, and *tch*. All words in which these graphemes appeared more than once were stored in a separate file, together with their frequencies to be processed later. With this provision, all trigrams and longer graphemes were treated as if they appeared only once in each word, simplifying the search algorithm. To see how the repeated graphemes above were treated, take *igh* for example. It appeared twice in 15 words, all derived from *highlight*. The sum of occurrences of these words was 155. Therefore, after searching for the *igh* grapheme, taken to appear only once in each word, another 155 occurrences had to be added to the number of occurrences found. Obviously, this simple manual proce-

dure was practical only for these multiletter graphemes, because they did not repeat in a large number of words. We used more sophisticated (and time-consuming) strategies later on, in order to search for repeated single-letter graphemes.

The search strategy used to count the frequency of bigrams and longer graphemes is now described briefly. It was implemented in the AWK programming language (Aho, Kernighan, & Weinberger, 1988), in a UNIX environment. The program developed takes as its input a grapheme and its corresponding phonemes, as given by Berndt et al. (1987). The grapheme *igh* and its corresponding phoneme “2” (/aɪ/) will be used as an example in what follows.

The program proceeds by scanning the CELEX database line by line, in search of a word containing the grapheme (*igh* in our example). When a word containing the grapheme *igh* is found, a trial search *width* is computed; (see Figure 1 for a graphical explanation of the search strategy). The search width is simply the difference between the number of characters in the word (length1 in Figure 1) and the number of characters in its phonemic transcription (in the CELEX character set), not counting diacritics (length2 in Figure 1). Thus, for the word *enlightenment* and its transcription “Inl2tHm@nt” (/ɪnlɑɪtɪnmənt/), which have 13 and 10 characters, respectively, the search width is 3. For many words, length1 and length2 are identical, in which case the search width is set to 1. Now, the position *k*, where the grapheme starts in the word is computed and one of the possible phonemes associated with it, is searched from *k* – width to *k* + width in the phonemic transcription. In our example, the grapheme *igh* appears at position *k* = 4, and width = 3. Thus, the *start* and *stop* positions to search for are *Start* = *k* – width = 1 and *Stop* = *k* + width = 7 (i.e., the search region goes from the 1st to the 7th character of

Figure 1. Searching for the trigram *igh* and its associated phoneme “2” in the word *enlightenment*. The search width and the start and stop positions for the search in the phonemic transcription are based on the position of the grapheme in the word (*k*) and the lengths of the word and its transcription.

the phonemic transcription of the word). The phonemic transcription of *igh*, the character “2” (/aɪ/), is then found at Position 4. Therefore, in this case, the search for *igh* pronounced as “2” (/aɪ/) succeeds without further complications. It is also apparent that, in this special case, the grapheme and the phoneme are both found as the 4th character in the word and in its transcription, respectively. The number of occurrences of the word *enlightenment* in the CELEX database is 45 times. Therefore, this number is added to the sum of occurrences of the grapheme *igh* pronounced as “2” (/aɪ/), and a new database line is read. If the grapheme *igh* occurs in it as well, the search algorithm above is repeated; if not, the next line is read. As seen before, in this example, the grapheme has only one phonemic transcription. For the majority of graphemes, which have multiple phoneme associations, each separate grapheme–phoneme association has its number of occurrences stored separately.

Some enhancements to the algorithm. More complex situations can occur, and three of them will be discussed next.

PECULIAR GRAPHEMES

Consider first the grapheme *mb*, which is pronounced as /m/ in words such as *climb* and *plumber*. There are, however words containing *mb*, such as *numbers* and *chambers*, in which *m* and *b* are both pronounced (i.e., they are two separate graphemes). Words such as these are excluded from the frequency count of the grapheme *mb*, by a simple additional constraint. When a word is found to contain the characters *mb*, the phonemic transcription is searched for the presence of the phoneme /m/ immediately followed by the phoneme /b/. If both are present, then the characters *m* and *b* are separate graphemes and are pronounced separately. In this case, the word is discarded, and a new line is read. Conversely, if the transcription contains /m/, but it is not immediately followed by /b/, then we have a genuine case in which *mb* is pronounced as /m/. In this case, the number of occurrences of the word is added to the variable storing the number of times *mb* was pronounced as /m/. This is an important enhancement to the algorithm, and it is extensively used when searching single-letter graphemes as explained below.

MULTIPLE PHONEMES FOUND FOR A SINGLE GRAPHEME

A more difficult problem occurs with graphemes that have many different pronunciations, such as *ai*, which can be pronounced in seven different ways (see Table 3): “2,” “8,” “@,” “E,” “{,” “I,” and “1” (/aɪ/, /eə/, /ə/, /ɛ/, /æ/, /ɪ/, and /eɪ/). In this case, the number of occurrences that *ai* is pronounced as “2,” *ai* is pronounced as “8,” “@,” and so on, have to be accumulated separately, as the database is scanned. There is no difficulty in that, and our program handles this correctly. A more subtle detail appears in words such as *traitors* and its transcription

Table 3
CELEX Search Results for Grapheme *ai*

Sum of Occurrences of Transcriptions		
Grapheme	Transcriptions	Occurrences
ai	2	359
ai	8	26,272
ai	@	3,394
ai	E	49,817
ai	{	178
ai	I	8,484
ai	1	46,581

“tr1t@z” (/tɹeɪtəz/) (see Figure 2). The word has eight characters, whereas the phonemic transcription has six, giving a search width of 2. The bigram *ai* starts at Position $k = 3$ in the word, so $\text{start} = k - \text{width} = 1$, and $\text{stop} = k + \text{width} = 5$. Thus, the search region goes from Character 1 (“t”) to Character 5 (“@”) in the phonemic transcription, and it is obvious that, in this case, two valid pronunciations for the bigram *ai* are found: “1” and “@” (/eɪ/ and /ə/). We know, however, that the bigram appears only once in the word. The second phoneme found, “@” (/ə/), corresponds to another grapheme (i.e., *or*). The solution to this problem is simple. Every time the number of pronunciations found for a bigram is larger than 1, the search width is reduced by one, and the search is repeated, as shown in Figure 2. If two or more pronunciations are still found, the width is reduced by 1 again, and a new search is performed. This procedure is repeated until only one pronunciation is found. In the example above, when the search width is reduced from 2 to 1, a single pronunciation is found. This word occurs 50 times in the corpus, and this number is added to the sum of the grapheme *ai* pronounced as “1” (/eɪ/).

GRAPHEMES WITHOUT ASSOCIATED PHONEMES

Under certain circumstances, the grapheme being searched is identified in a word; however, for some reason, its corresponding phoneme is not found in the transcription. In this case, the whole line containing the word, its frequency of occurrence, and transcription is printed on the screen and also sent to a file for further manual processing. This situation can occur for many reasons, the most common being the case in which a new phonemic transcription (not included in Berndt et al., 1987) is associated with the grapheme under search. This was the case for the grapheme *ai* used in the examples above. In Berndt et al., there are only five phonemes associated with it: “1,” “E,” “I,” “@,” and “{” (/eɪ/, /ɛ/, /ɪ/, /ə/, and /æ/). In the present algorithm, we found that, in the CELEX database, it has two more possible pronunciations: “2” and “8” (/aɪ/ and /eə/). The program was then run a second time for the *ai* grapheme, including the two new pronunciations, and

Figure 2. Searching for the bigram *ai* and one of its associated phonemes ("2," "8," "@," "E," "{," "I," or "1") in the word *traitors* and its phonemic transcription "tr1t@z" /treɪtəz/, respectively. k is the position where the grapheme starts in the word. The search region, from the start character to the stop character, has to be reduced by reducing width, in order to find a unique phoneme associated with this grapheme.

the frequencies of occurrence of these grapheme–phoneme associations were computed. A few obviously wrong transcriptions in the CELEX database were also uncovered by this method, which forced problematic words to be printed on screen.

Another case that produced this output was the character *h* at the beginning of words such as *hour*. This is a genuine case in which the character *h* is not pronounced (i.e., it has no corresponding phoneme).

When the program reaches the end of the CELEX database, it will have counted the number of times the given grapheme was pronounced as each of the given phonemes. The program then produces a table containing in each line the grapheme, one of its associated phonemes, and the number of occurrences found (as exemplified in Table 3 (for the grapheme *ai*). This table is saved to a file, and subsequent runs of the program, for new graphemes, append their tables to this file. Thus, a new database containing grapheme–phoneme associations and their number of occurrences was generated. In the next section, we discuss the search strategy for single-letter graphemes.

Searching Single-Letter Graphemes

The search strategies used in this case are similar to the ones described above. The task is made manageable by the

fact that the multiletter graphemes have already been searched. Thus, our list of multiletter graphemes provides us with a guide to check whether a single letter in a word is a grapheme on its own or whether it is part of a multiletter grapheme. We essentially use the same procedure implemented before for peculiar graphemes, to decide whether two letters formed a bigram or two separate graphemes. As an example, in some words, the two consecutive characters *ai* will form two separate graphemes (i.e., we will find in the phonemic transcription a phoneme corresponding to *a*, immediately followed by another one, corresponding to *i*). In such cases, *a* and *i* are two separate graphemes. There are other words, however, whose phonemic transcription contains only one phoneme associated with *ai* as seen before.

Another aspect of single-letter graphemes is that they can repeat any number of times in a word. This is a further complication to be dealt with in the search strategy. Once a word is found to contain the grapheme under search, a check is made to determine how many times it appears in the word. If it occurs only once, the search proceeds as in the previous section. If it appears n times ($n \geq 2$), all positions it occupies in the word are stored. Essentially, the integer k used in the previous section to store the position of the bigram in a word will now be an array of integers ($k_1, \dots, k_n$), storing all n positions where the grapheme is found. A search is

triggered starting at k_1 – width, for the phoneme corresponding to the first occurrence of the grapheme. This procedure will be repeated n times, until all n grapheme–phoneme correspondences in the word are accounted for.

In the process of searching the database, many new grapheme–phoneme associations, not present in Berndt et al. (1987), came to light. All such cases have been included in the database in the Appendix. In all, 461 grapheme–phoneme associations were found, with 195 unique graphemes, giving 2.4 phonemes associated with each grapheme, on average.

Computing Probabilities

Having obtained the grapheme–phoneme associations and the number of occurrences for each of them, it is now possible to calculate the percentage of occurrences of each grapheme in the corpus. First, all occurrences of all graphemes were added, producing a grand total of 67,590,620. Dividing the total number of occurrences of a grapheme in the corpus (i.e., including all phonemes associated with it) by the grand total above ($\times 100\%$) gives the percentage of occurrence of that grapheme in the corpus. This ratio of the grapheme total and the grand total can also be viewed as a good estimate for the prior probability of occurrence of the given grapheme in the (British) English language. If the grapheme a is taken as an example, its total number of occurrences is obtained by adding the number of occurrences for all its associated phonemes. This gives a total of 3,746,713, which divided by the grand total gives 0.055. Therefore, the prior probability of appearance of the grapheme a (in all its phonemic associations) in the corpus is 5.5%. It should be noted that this is a reasonably high percentage, because it does not include combinations of a and other characters, which are independent graphemes.

As the number of occurrences of each separate grapheme–phoneme association has also been computed, the percentages of occurrences for each individual grapheme–phoneme association can be obtained. For example, a pronounced as “@” (/ə/) occurred 591,123 times in the CELEX corpus. Dividing this by the total number of occurrences of grapheme a (3,746,713) gives the percentage of occurrences of a being pronounced as “@” (/ə/). The number obtained is .158, which can be interpreted as the probability ($\times 100\%$) of a being pronounced as “@” (/ə/). The probabilities for the remaining phonemes associated with a have been calculated in the same manner and can be seen in the Appendix. Similar calculations were performed for all graphemes and for all grapheme–phoneme associations, producing the corresponding grapheme and grapheme–phoneme probabilities. In the Appendix, all grapheme–phoneme associations are listed, together with both probabilities, for the grapheme as a whole (prior probability) and for each of its phonemic associations. Also included for each grapheme–phoneme association in the Appendix is an example word and its phonemic transcription.

Database Validation and Consistency Checks

After obtaining all probabilities, simple tests were run, to ensure consistency and to check the database for any mistakes. It is clear from the discussion above, that the grapheme prior probabilities are normalized to 1 (i.e., the sum of all grapheme prior probabilities P_{gi}^p adds to 1):

$$\sum_i P_{gi}^p = 1, \quad (1)$$

where P_{gi}^p is the total (prior) probability of appearance of grapheme g_i in the database.

The grapheme–phoneme probabilities P_{gi}^p , taken for each grapheme separately, are also normalized to 1:

$$\sum_{i=1}^k P_{gi} = 1, \quad (2)$$

where k equals the number of phonemes associated with grapheme g_i . Short AWK programs were written to perform these sums on the database presented in the Appendix, and both types of probabilities were found to be correctly normalized.

Because the basic global properties of the database agree with Equations 1 and 2, we proceeded to validate the individual statistical information for the grapheme–phoneme associations. Perhaps the best test possible for the consistency of the grapheme–phoneme statistical database would be to use it to predict the pronunciation of real words taken from the CELEX database. This would check the accuracy of the statistical information obtained against some of the words (and their phonemic transcriptions) used for its derivation. Strictly speaking, actual pronunciation cannot be predicted solely on the basis of the grapheme–phoneme associations in the Appendix, as stress information is not available there. Still, it should serve to retrieve the correct phonemes contained in a given word.

A final simple check (suggested by one of the reviewers of this work) was performed, to test whether the search procedure led to any wrong association. It was seen before that when all occurrences of all graphemes were added, a grand total of 67,590,620 was obtained. The same number should be obtained by multiplying, for each word, its frequency by its number of phonemes and by adding across all words. The result of this calculation was 67,583,526, being lower than the grand total by 7,094 grapheme–phoneme correspondences (i.e., 7,094 correspondences were counted twice by our search algorithm). This corresponds to only 0.005% of the total. Therefore, some correspondences in the Appendix are higher than in reality, by a tiny amount. Clearly, we cannot claim our correspondences to be totally error-free, but they are remarkably close to that.

A number of different schemes can be employed to use the statistical information obtained before, in order to predict word pronunciation. Metrics based on the present statistics, coupled with rules, perhaps similar to those devised by Venezky (1970), could prove very effective in pronun-

ciation prediction. However, only the simple metric proposed in Berndt et al. (1987) was employed to test the database statistical content and to provide a simple means of comparison with previous work. Let $P(p|g)$ be the probability of grapheme g being pronounced as phoneme “p,” and let P_g^* be the probability of the most frequent grapheme–phoneme correspondence for the grapheme g . For instance, $P_A^* = .496$, and $P(\text{ə}|A) = .1578$, $P(\text{æ}|A) = .496$, and so on (see Appendix). Consider a printed word to be given by a sequence of graphemes (containing one or more characters each) $g_1, g_2, \dots, g_n$. Its correct pronunciation would then be given by a string of phonemes $p_1, p_2, \dots, p_n$. An overall metric of predictability of the pronunciation of such a word is given by Berndt et al. (1987):

$$m = \frac{1}{n} \sum_{i=1}^n \frac{P(p_i|g_i)}{P_{g_i}^*}. \quad (3)$$

Clearly, $P(p_i|g_i)/P_{g_i}^*$ is simply the probability of grapheme g_i being pronounced as “p_i” divided by the most probable pronunciation of grapheme g_i , $P_{g_i}^*$. Therefore, Equation 3 above is simply the average of these ratios, for all the n graphemes in the word. This metric of predictability was used in Berndt et al. (1987) to calculate the probability with which the statistical grapheme–phoneme correspondences would predict correctly the pronunciation of a given word. In the present work, we went one step further and used Equation 3 to obtain the probabilities of all the possible pronunciations of a given word. For example, it can be seen in the Appendix that there are three, nine, and one phonemes associated with the graphemes t , a , and p , respectively. Therefore, the number of possible pronunciations of the word *tap* will be $3 \times 9 \times 1 = 27$. Let t_i be the number of phonemes “p_i” associated with grapheme g_i . In general, the number of possible pronunciations “p₁,” “p₂,” ..., “p_n” of a word given by the sequence of graphemes $g_1, g_2, \dots, g_n$ will be $t_1 \times t_2 \times \dots \times t_n$. Obviously, the correct pronunciation will not always be the one having the highest ($m = 1$) probability. Using the same examples given in Berndt et al. (1987) for comparison, our database produces for the word *tap*:

$$\begin{array}{lll} t \rightarrow t & [P(t|t) = .940 & P_t^* = .940] \\ a \rightarrow \text{æ} & [P(\text{æ}|a) = .496 & P_a^* = .496] \\ p \rightarrow p & [P(p|p) = 1 & P_p^* = 1]. \end{array}$$

Therefore,

$$m = \frac{1}{3} \left(\frac{.940}{.940} + \frac{.496}{.496} + \frac{1}{1} \right) = 1.$$

Thus, the statistical information included in the Appendix produces the correct pronunciation, in this case. However, as seen above, the word *tap* has 27 possible pronunciations associated with it. Despite this relatively large number of possible pronunciations, the correct pronunciation is the one producing the highest value of m . In conclusion, it can be said that the pronunciation of the word *tap* is highly predictable from the grapheme–phoneme associations of its component graphemes.

For another example (the word *who*), in Berndt et al. (1987):

$$\begin{array}{lll} wh \rightarrow h & [P(h|wh) = .204 & P_{wh}^* = .796] \\ o \rightarrow u & [P(u|o) = .162 & P_o^* = .363] \end{array}$$

$$m = \frac{1}{2} \left(\frac{.204}{.796} + \frac{.162}{.363} \right) = .35.$$

Two things should be noticed in this case. First, the pronunciation with the highest m ($m = 1$) is “wɒ” (/wɒ/), but this is not the correct pronunciation. There are in this case 18 (9×2) pronunciations, and, if they are sorted in decreasing order of m , the correct pronunciation /hu/ (/hu:/) occupies the 12th position. Second, $m = .35$ for the correct pronunciation, as opposed to $m = .09$, given in Berndt et al. (1987). This discrepancy can be explained by the vast size of the corpus on which the present statistical information is based. This should be more representative of the frequency of occurrence of graphemes in (British) English and produced a higher probability of “o” being pronounced as “u” and also of *wh* being pronounced as “h.” Therefore, even if the correct pronunciation is not the most probable, it could be argued that this value of m is more accurate than that found in Berndt et al. (1987). Especially for low-frequency grapheme–phoneme correspondences, the size of the corpus used can have a significant effect on the probabilities obtained. Also, the fact that American and British English were used in the previous study and in the present work respectively will undoubtedly change the probabilities of certain grapheme–phoneme correspondences.

PRONUNCIATION PREDICTION

It would be extremely tedious to compute manually all the possible pronunciations of any given word (or, more generally, any string of graphemes) and their associated probabilities. This is obviously a task best performed by a computer program. Accordingly, a program was developed that, given a word already divided in graphemes, searches the statistical database in the Appendix and computes all possible pronunciations of the word and their associated probabilities. The task of dividing a given word in its associated graphemes has not been tackled in this work and has to be performed by the user. The probabilities are calculated by Equation 3, and they will be referred to as “ m -scores” from now on. Before displaying the results, they are sorted in decreasing order of m -scores, and the number of possible pronunciations to be displayed can also be controlled by the user.

As we have discussed previously, the deep orthography of English allows the existence of two kinds of words: regular words, which follow the rules of spelling-to-sound correspondence, and exception words, which break them. There is a large number of psycholinguistic studies devoted to the responses to these two kinds of words. It has been found, for example, that regular words take less time to be named than do exception words (e.g., Glushko, 1979; Gough & Cosky, 1977). An obvious use for the probabilities obtained in the present work is the preparation of ex-

perimental materials to test regularity effects in the English lexicon. Furthermore, it can assist in the prediction of how regular the pronunciation of a pseudoword or a nonword would be, according to the lexicon. For example, a group of trained phoneticians were presented with nonwords and asked to give all their possible phonemic transcriptions according to the rules of English (Gontijo, 1998). A high correlation was found between the number of times each pronunciation was given by the phoneticians and its m -score as calculated from Equation 3.

Using the m -score program, it is possible to probe many interesting aspects of the statistical information contained in the Appendix. It was found that certain words, despite being long and containing many vowels, have highly predictable pronunciation. One such example is the word *understand*, which has nine graphemes (*u-n-d-e-r-s-t-a-n-d*) and, because they have many different pronunciations, produce a vast number of possible combinations. The total number of possible pronunciations for this word is 209,952 ($=9 \times 3 \times 2 \times 6 \times 4 \times 3 \times 9 \times 3 \times 2$; see the grapheme-phoneme correspondences for the nine graphemes in the Appendix), and it is remarkable that the program finds the correct pronunciation [i.e., $m = 1$ for "Vn-d@-st{nd}" ($\wedge n-d\text{-}st\text{-}\text{end}/$)] as being the most probable, solely on the basis of the simple metric of predictability described by Equation 3. On the other hand, there are deceptively simple words, such as *examine*, for which the statistical information alone cannot be used even as a first approximation to the correct pronunciation. This word has seven single-letter graphemes and 2,592 possible pronunciations; the correct one is given as the 805th by the program. Even in this case, though, the failure of the metric used is not as bad as it might appear. The value of m obtained in this case was $m = .83$, which is still very close to the most probable, $m = 1$. In this case, the program fails to detect the correct pronunciation with any accuracy, but it

still gives important information about the word. There is a large number (804) of possible phoneme combinations (pronunciations), all having very close values of m , between $m = 1$ (the most probable) and $m = .83$ (for the correct pronunciation). This type of "phonemic neighborhood" information can be useful to guide the choice of material for experiments on phonology, for example.

The program above was also used to predict the pronunciation of words taken from the CELEX database. The results produced could then be compared with the actual pronunciations, also listed in CELEX. Obviously, the frequency of occurrence of a word will have a big influence on the predictability of its pronunciation. In a high-frequency word, at least some of its graphemes will also have a high frequency of occurrence. Thus, the correct pronunciation of high-frequency words should be easier to predict. Five sets of 100 words each, chosen at random, were extracted from CELEX for the tests. The Sets 1, 2, 3, 4, and 5 contain words whose frequencies of occurrence are larger than or equal to 0, 100, 1,000, 5,000, and 10,000, respectively. Therefore, words in Set 1 are chosen completely at random from the CELEX database. The others are still chosen at random, but with the provision that their frequencies are above the threshold set for each group (100, 1,000, and 10,000).

The program was run for each individual word in each group, and the position of the correct pronunciation (i.e., if it was the 1st, 2nd, 3rd, 4th, and so on, in decreasing order of m) was saved. Figure 3 shows the results of this test graphically, where only the first 20 most probable pronunciations were taken into consideration. Thus, if the pronunciation produced by the program is at a position beyond the 20th, it is considered that the program failed to detect it. For each set, the y-axis represents the number of words whose pronunciations can be predicted by the first, first 2, ..., first 20 pronunciations produced by the pro-

Figure 3. Predicting the pronunciation of words chosen at random from CELEX. The higher the frequency of a word, the easier it is to predict its pronunciation. The lines connecting the points are guides to the eye only.

gram. Therefore, if the program is “allowed only one guess,” it can predict the pronunciation of 24, 32, 38, 38, 48 words for the Sets 1, 2, 3, 4, 5, respectively (see Figure 3, y-axis). This general trend of higher predictability with higher frequency of occurrence of the words is maintained throughout the sets. Thus, this test can be considered as a validation of the grapheme–phoneme correspondences extracted from CELEX.

During the testing described above, it was noticed that the shorter the word, the easier it was, in general, to predict its pronunciation. This was not surprising, since the majority of words in the corpus from which the probability database was derived had fewer than four phonemic syllables (see Table 2). Also, shorter words have fewer phonemes, resulting in fewer possible pronunciations. This motivated one further test, in which only monosyllables were used. Again, five sets of 100 monosyllables each, with frequencies larger than 0, 100, 1,000, 5,000, and 10,000 were chosen at random from the CELEX database. For the 5,000 and 10,000 sets, there were only 79 and 63 monosyllables, respectively, in the corpus. So, the results for these two sets were normalized to 100 words. As before, the program was run for each individual word, and the results are shown in Figure 4, where the axes have the same meaning as in Figure 3. The first thing to notice is that the five curves are much closer together here than in Figure 3. Therefore, the frequency of the monosyllables in each set is less important for pronunciation prediction than it is for words of uncontrolled length (in Figure 3).

As expected, the monosyllables have pronunciations easier to predict, compared with the words used in Figure 3. For example, if a level of five pronunciations (five allowed guesses) is chosen, then the program can predict the pronunciation of 51, 69, 70, 79, and 84 words in the five sets of words used in Figure 3. For the monosyllables (Figure 4), the corresponding figures are 76, 76, 84, 85,

and 78. Because there are not enough monosyllables with frequencies above 10,000, the statistics for this set of words are not as reliable as for the others. This could explain the fact that fewer words were correctly predicted for this set in five guesses than for the 1,000 and 5,000 sets.

The extensive testing presented in Figures 3 and 4 demonstrate that probability alone is not enough to produce high levels of accuracy in pronunciation prediction. However, if the goal is accurate pronunciation prediction, then the statistical information obtained here can be combined with rules, perhaps similar to those of Venezky (1970), to produce a better metric for measuring the pronunciation predictability. On the other hand, the type of statistical information presented here is interesting in itself and might be useful to control the “frequency variable” in experiments involving the pronunciation of strings of graphemes (nonwords and pseudowords) not found in the (British) English lexicon.

Another use of this statistical information is to calculate the “phonemic neighborhood” of a given string of graphemes, for example. In this context, it is interesting to recall the two example words discussed before: *understand* and *examine*. When applied by those using nonwords and pseudowords in pronunciation experiments, grapheme–phoneme statistics can become an important tool for the selection of materials.

A METHOD OF MEASURING THE ORTHOGRAPHIC DEPTH OF A LANGUAGE

The relationship between a script and its language is described by orthography (Scheerer, 1986). The variety of features that exist in spoken languages has given rise to many orthographies, each reflecting a unique relationship between the structure of a language and its script. During

Figure 4. Predicting the pronunciation of monosyllables chosen at random from CELEX. The frequency of occurrence has less effect here than in Figure 1. Again, the lines are guides to the eye only.

the development of a language, the attempt to make an efficient match between the written form and its morphology and phonology will determine whether the resulting orthography will be a syllabary, a syllabary-cum-logography, or an alphabet. Alphabetic orthographies are those in which the letters composing the script are matched to the language phonemes. Languages with alphabetic orthographies present varying degrees of dependence on the strict alphabetic principle: The range of correspondence between grapheme and phoneme varies in both consistency and completeness. The reason for the differences is largely accounted for by the particular phonological and morphological characteristics of each language.

As discussed in the introduction, languages have either a "shallow" or a "deep" orthography depending on the ease of predicting the pronunciation of a word from its spelling. In shallow orthographies, the spelling-sound correspondence is direct: Given the rules, anyone can immediately name words correctly. Shallow orthography languages are not phonologically complex. All morphologically related words have a common word stem that will necessarily sound the same in all of them. An extreme example of this kind is the Serbo-Croatian language, which follows the principle, "Spell a word like it sounds and speak it the way it is spelled" (Katz & Frost, 1992, p. 69). In contrast, in deep orthographies, the relationship is less direct, and readers must learn the arbitrary or unusual pronunciations of irregular words. The English language is an example of a deep orthography language, in which there is a large amount of phonologic change among words from a single derivational family. Examples of this are the phonological variations in *heal* and *health* and in *steal* and *stealth*.

Research on the orthographic depth of languages and on how orthography influences visual word recognition is important because it has the potential to reveal both universal and orthography-specific aspects of reading (Seidenberg, 1992). A question of interest is whether this deep/shallow dimension of orthography leads to differences in the recognition of printed words. For example, how different is reading in Spanish, which conforms closely to the alphabetic principle, from reading in modern Hebrew, which does not bear a one-to-one relationship with its nonvocalized orthography (Ravid, 2001)?

The orthographic depth hypothesis (ODH) has been put forward, which states that the more transparent (or shallow) an orthography is, the more efficiently phonological representations will mediate between print and lexicon (Katz & Frost, 1992). Such a phonological representation is assembled by the reader's brain as he/she uses the orthography correspondences between subword spelling and sound. In contrast, (deep) opaque orthographies encourage the reader to process the printed words by referring to their printed morphology, as opposed to their phonology. In other words, the ODH asserts that oral reading in shallow orthographies is qualitatively different from oral reading in deep orthographies. This hypothesis is not universally accepted, however, and the reader is referred to Katz and Frost for an enlightening treatment of this subject.

Despite a great deal of cross-linguistic research, there does not seem to exist a quantitative way of measuring the orthographic depth of different languages. It is normally assumed that Spanish and Portuguese have shallow orthographies, whereas English, unpointed Hebrew, and German are examples of deep orthographies. However, it is not clear how large a difference in orthographic depth there is, on one hand, between Spanish and Portuguese and, on the other hand, between English, unpointed Hebrew, and German. To put this in a context more relevant to educators, how different are the reading processes used by native speakers of these languages? Scheerer (1986) has shown that native speakers of German use grapheme-phoneme conversion rules both in pronunciation tasks and in lexical decision tasks. Many researchers who focus primarily on English have traditionally denied the involvement of such rules in reading (Scheerer, 1986, p. 268). Therefore, even among languages normally clustered together as deep orthographies, there are variations in the reading process. We suggest that a quantitative procedure to measure and compare the orthographic depth of languages could be very useful in this context. Perhaps the numerical value of orthographic depths of languages could become a good way of estimating whether or not reading in these languages follow similar processes. This obviously would improve our understanding of knowledge representation and retrieval by the brain.

Thus, one of the aims of the present study was to propose a general way of measuring orthographic depth of languages. The statistical information presented in the Appendix can be used to estimate the orthographic depth of English, as explained below. If the same type of statistical information is obtained for other languages, then their orthographic depth could be calculated as well, and many cross-linguistic comparisons could be made.

A measure of the orthographic depth of a language can be calculated by taking the total number of grapheme-phoneme correspondences (GPC) in the language and dividing it by the total number (g) of graphemes in the language. Therefore, we are postulating that the ratio (GPC/g)—that is, the average number of phonemes corresponding to a grapheme of the language—is a direct measurement of its orthographic depth. In the case of English, for example, we have seen that the total number of graphemes (g) found in the CELEX database was 195, and the total number of grapheme-phoneme correspondences amounted to 461. Thus, the orthographic depth of English is calculated to be $GPC/g = 2.4$.

A (small and incomplete) list of Brazilian Portuguese grapheme-phoneme correspondences can be found in the *Collins Gem Dictionary* (Harland, 1987). It contains the 57 most frequent graphemes, and it is interesting that the number of grapheme-phoneme correspondences for this list is only 61. This gives Portuguese an orthographic depth value of 1.07, which should be contrasted with the value obtained above for English. It can be argued that the comparison above is not entirely valid, because the sizes of the corpora used were different. In order to make a more accurate comparison, one should take only the 57

most frequent English graphemes (g) from the Appendix. These graphemes have 223 phonemes (GPC) associated with them. Thus, the orthographic depth ratio for English in this case is $GPC/g = 3.9$, which obviously strengthens the difference in the depth of English and Portuguese.

Another possible way of measuring the orthographic depth of a language can be derived from the present work. In order to implement it, a large number of high-frequency words have to be randomly selected. The program described above then calculates their possible pronunciations and m -scores. The words are then checked to determine whether the program correctly predicted their pronunciation (i.e., if $m = 1$ for the correct pronunciation). The number of words for which this is true is divided by the total number of words searched and multiplied by 100, to give a percentage. It is this percentage that gives another measure of the orthographic depth of the language. For example, in Figure 3, this percentage is found to be 48% for the group of words with frequency above 10,000. As mentioned before, this is actually a low value, resulting from the deep orthography exhibited by British English. At present, we have no means of computing these percentages for other languages. However, we predict that this percentage would be much higher for Brazilian Portuguese (i.e., it would be easier to predict the correct pronunciation of words in Brazilian Portuguese).

Finally, it can be seen that the two methods of measuring orthographic depth proposed here are equivalent. In a hypothetical language in which each grapheme has only one corresponding phoneme, the first method would produce $GPC/g = 1$. This is actually the number to be obtained for the language with the shallowest possible orthography, in which there is a one-to-one correspondence between graphemes and phonemes. At the same time, there would be only one possible pronunciation for each word in such a language (see Equation 3 and the discussion following it). Therefore, the m -scores would always be equal to 1, and the resulting percentage of words whose pronunciations can be predicted correctly would be 100%. Again, this would be a measure of the shallowest possible orthography, according to our second method of depth measurement.

On the other hand, for a hypothetical language in which each grapheme corresponds to, say, 10 possible phonemes, the $GPC/g = 10$. This would be a language of very deep orthography. In addition, even for short words, composed, for example, of three graphemes only, there would be 1,000 possible pronunciations. It would be unlikely that the pronunciation with $m = 1$ would be the correct one for a large number of words. Therefore, the percentage of words whose pronunciation could be predicted correctly would be small. Again, we would conclude that such a language would have a deep orthography.

CONCLUSIONS

The frequencies of occurrence of 461 grapheme-phoneme correspondences in British English have been

obtained through an exhaustive search of the CELEX database. These results, included in the Appendix, can be very useful in studies such as spelling and speech synthesis. In addition, it was shown that these frequencies could be used to predict all the possible pronunciations of any given string of graphemes. The computer program developed to accomplish this task showed that the correct pronunciation of very high-frequency words can be predicted with a high degree of accuracy. The prediction power of this procedure decreases in general for long or low-frequency words, but it could be improved easily by complementing its statistical algorithm with a small set of rules.

All the possible pronunciations of a nonword, together with their probabilities, can also be obtained by this method. This could become a useful tool in the preparation of materials for psycholinguistic experiments.

Finally, all the results obtained in this study led to a quantitative method to measure the orthographic depth of different languages. In particular, Brazilian Portuguese was compared with British English and shown to have a much shallower orthography.

REFERENCES

- AHO, A. V., KERNIGHAN, B. W., & WEINBERGER, P. J. (1988). *The AWK programming language*. New York: Addison-Wesley.
- BAAAYEN, R. H., PIEPENBROCK, R., & GULIKERS, L. (1995). *The CELEX lexical database* (Release 2, English Version 2.5) [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.
- BERNDT, R. S., D'AUTRECHY, C. L., & REGGIA, J. A. (1994). Functional pronunciation units in English words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 977-991.
- BERNDT, R. S., REGGIA, J. A., & MITCHUM, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments, & Computers*, **19**, 1-9.
- GLUSHKO, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 674-691.
- GONTIJO, P. F. D. (1998). *Familiarity effects in visual word recognition*. PhD thesis, University of Edinburgh, Scotland.
- GOUGH, P., & COSKY, M. (1977). One second of reading again. In N. Castellan, D. Pisoni, & G. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 271-288). Hillsdale, NJ: Erlbaum.
- HANNA, J. S., & HANNA, P. R. (1959). Spelling as school subject: A brief history. *National Elementary Principal*, **38**, 8-23.
- 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. Government Printing Office.
- HARLAND, M. (1987). *Collins Gem Portuguese dictionary*. London: Collins.
- HENDERSON, L. (1982). *Orthography and word recognition in reading*. London: Academic Press.
- KATZ, L., & FELDMAN, L. B. (1981). Linguistic coding in word recognition: Comparisons between a deep and shallow orthography. In A. M. Lesgold & C. A. Perfetti (Eds.), *Interactive processing in reading* (pp. 85-106). Hillsdale, NJ: Erlbaum.
- KATZ, L., & FROST, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 67-84). New York: North-Holland.
- RAVID, D. (2001). Learning to spell in Hebrew: Phonological and morphological factors. *Reading and Writing: An Interdisciplinary Journal*, **14**, 459-485.
- SCHERER, E. (1986). Orthography and lexical access. In G. Augst (Ed.), *New trends in graphemics and orthography* (pp. 262-286). Berlin: De Gruyter.

- SEIDENBERG, M. S. (1992). Beyond orthographic depth in reading: Equitable division of labor. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 85-118). New York: North-Holland.
- TREIMAN, R., BIJELJAC-BABIC, R., & RICHMOND-WELTY, E. D. (1995). The special role of rimes in the description, use and acquisition of English orthography. *Journal of Experimental Psychology: General*, **124**, 107-136.
- VENEZKY, R. L. (1970). *The structure of English orthography*. The Hague: Mouton.
- WIJIK, A. (1966). *Rules of pronunciation for the English language*. Oxford: Oxford University Press.
- ZIEGLER, J. C., STONE, G. O., & JACOBS, A. M. (1997). What is the pronunciation for -ough and the spelling for /u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments, & Computers*, **29**, 600-618.

APPENDIX **Grapheme-Phoneme Correspondences in British English**

PP is the probability of occurrence of the grapheme in the lexicon.

Prob is the probability of the grapheme being pronounced as the corresponding phoneme.

Word shows an instance of the grapheme-phoneme correspondence.

Pronunciation gives the pronunciation of the example word, using the CELEX "disk" character set.

NP = not pronounced.

a#e = a followed by another grapheme, with e as the last character in a word, such as in *face*.

The probabilities below are shown up to four decimal places only. Thus, the grapheme-phoneme correspondence "a#e - q" is shown with $p = .0000$, but the real value is $p = .00004120$. The corresponding author can be contacted regarding electronic copies of the software and of this appendix, with eight decimal places.

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
a	.0554	{	.4964	had	'h{d
a		@	.1578	about	@-'b6t
a		#	.0885	after	'#f-t@R
a		1	.0797	taken	't1-k@n
a		Q	.0768	was	'wQz
a		\$	.0651	all	'\$l
a		E	.0140	many	'mE-nI
a		I	.0114	manager	'm{-nI-@R
a		8	.0102	various	'v8-r7s
a#e	.0097	1	.3623	face	'f1s
a#e		{	.2255	collapse	k@-'l{ps
a#e		@	.1735	ultimate	'V1-tI-m@t
a#e		#	.1699	morale	mQ-'r#l
a#e		I	.0395	heritage	'hE-rI-tI_
a#e		8	.0216	square	'skw8R
a#e		\$	.0074	false	'f\$ls
a#e		U	.0001	immaculate	I-'m{-kjU-IUt
a#e		Q	.0001	blancmange	bl@-'mQn_
a#e		q	.0000	melange	m1-'lqnZ
ae	.0001	2	.2408	maestro	'm2s-tr5
ae		@	.1923	gynaecologist	'g2-n@-'kQ-l@-Ist
ae		1	.1868	Israeli	Iz-'r1-lI
ae		i	.1648	aesthetic	is-'TE-tIk
ae		I	.1053	anaesthesia	'{-nIs-'Ti-zj@
ae		8	.0655	aeroplane	'8-r@-p11n
ae		7	.0365	Israelite	'Iz-'r7-l2t
ae		E	.0080	haemorrhage	'hE-m@-rI_
ah	.0001	@	.5510	Messiah	mI-'s2-@
ah		#	.4381	Brahmin	'br#-mIn
ah		1	.0056	dahlia	'd1-lj@z
ah		{	.0053	Fahrenheit	'f{-r@n-h2t
ai	.0020	E	.3688	again	@-'gEn
ai		1	.3448	detail	'di-t1l
ai		8	.1945	dairy	'd8-rI
ai		I	.0628	captain	'k{p-tIn
ai		@	.0251	chieftain	'Jif-t@n
ai		2	.0027	Kaiser	'k2-s@R
ai		{	.0013	plaid	'pl{d

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
ai#e	.0001	1	.8029	appraise	@-'pr1z
ai#e		8	.1466	questionnaire	'kwEs-J@-'n8R
ai#e		2	.0499	aisle	'2l
ai#e		E	.0005	bouillabaisse	'bu-j@-bEs
aigh	.0000	1	1.0000	straight	'str1t
air	.0002	8	1.0000	airline	'8-12n
al	.0022	P	.8761	capital	'k{-pI-tP
al		\$	.1239	talking	't\$-kIN
ao	.0000	\$	.7921	extraordinary	Iks-'tr\$-dH-rI
ao		6	.2079	Maoist	'm6-1st
ar	.0048	#	.7775	cards	'k#dz
ar		@	.1352	afterwards	'#f-t@-w@dz
ar		\$	.0829	awards	@-'w\$dz
ar		8	.0044	stares	'st8z
ar#e	.0001	#	1.0000	charge	'J#_
are	.0001	8	1.0000	awareness	@-'w8-nIs
au	.0004	\$	.7172	authority	\$-'TQ-r@-tI
au		#	.1929	aunt	'#nt
au		@	.0646	restaurant	'rE-st@-r~N
au		Q	.0111	sausages	'sQ-sI-_Iz
au		5	.0109	chauffeur	'S5-f@R
au		6	.0033	Frau	'fr6
au#e	.0004	Q	.8000	because	bI-'kQz
au#e		\$	.1928	cause	'k\$z
au#e		5	.0049	mauve	'm5v
au#e		#	.0024	auntie	'#n-tI
augh	.0001	\$	1.0000	daughter	'd\$-t@R
aw	.0004	\$	1.0000	raw	'r\$
aw#e	.0000	\$	1.0000	awe	'\$
ay	.0027	1	.9097	day	'd1
ay		I	.0830	yesterday	'jE-st@-dI
ay		8	.0047	prayer	'pr8R
ay		2	.0027	Mayan	'm2-@n
ay#e	.0000	1	.9168	aye	'1
ay#e		2	.0832	aye	'2
b	.0171	b	1.0000	back	'b{k
bb	.0001	b	1.0000	rubble	'rV-bP
bt	.0001	t	1.0000	subtle	'sV-tP
c	.0230	k	.6742	case	'k1s
c		s	.2953	cent	'sEnt
c		S	.0300	depreciate	dI-'pri-SI-1t
c		J	.0005	cello	'JE-l5
cc	.0005	k	1.0000	accurate	'{-kjU-r@t

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
cch	.0000	k	1.0000	saccharine	's{-k@-r2n
ce	.0000	S	1.0000	curvaceous	k3-'v1-S@s
ch	.0043	J	.8694	each	'iJ
ch		k	.1045	character	'k{-r@k-t@R
ch		S	.0204	machine	m@-'Sin
ch		—	.0057	sandwich	's{-n-wI_
cht	.0000	t	1.0000	yacht	'jQt
ci	.0007	S	1.0000	social	's5-SP
ck	.0016	k	1.0000	back	'b{k
cq	.0000	k	1.0000	acquaintance	@-'kw1n-t@ns
ct	.0000	t	1.0000	indictment	In-'d2t-m@nt
cz	.0000	z	1.0000	czar	'z#R
d	.0417	d	.9947	day	'd1
d		—	.0053	education	'E-_U-'k1-SH
dd	.0004	d	1.0000	middle	'mI-dP
dg	.0003	—	1.0000	bridge	'brI_
di	.0000	—	1.0000	soldiers	's5l-_@z
dj	.0000	—	1.0000	adjustment	@-'_Vst-m@nt
e	.0485	E	.3369	get	'gEt
e		@	.3017	after	'#f-t@R
e		I	.2600	because	bI-'kQz
e		i	.0349	female	'fi-m1l
e		j	.0290	azaleas	@-'z1-lj@z
e		3	.0233	her	'h3R
e		8	.0072	parent	'p8-r@nt
e		1	.0054	elite	l-'lit
e		#	.0016	clerk	'kl#k
e#e	.0335	i	.7382	complete	k@m-'plit
e#e		I	.0753	college	'kQ-Il_
e#e		E	.0507	fence	'fEns
e#e		@	.0501	dependence	dI-'pEn-d@ns
e#e		8	.0419	where	'w8R
e#e		3	.0280	reserve	rI-'z3v
e#e		7	.0097	severe	sI-'v7R
e#e		1	.0062	creche	'kr1S
ea	.0069	i	.3610	beach	'biJ
ea		7	.2223	real	'r7l
ea		E	.1654	ahead	@-'hEd
ea		I	.1230	forehead	'fQ-rId
ea		1	.0521	greatest	'gr1-t1st
ea		@	.0468	sergeant	's#-_@nt
ea		8	.0181	bear	'b8R
ea		#	.0092	thereafter	'D8-'r#f-t@R
ea		{	.0020	whereas	w8-'r{z

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
ea#e	.0006	i	.5026	leave	'liv
ea#e		@	.2406	vengeance	'vEn-@ns
ea#e		7	.1168	laureate	'l\$-r7t
ea#e		E	.0790	cleanse	'klEnz
ea#e		I	.0610	mileage	'm2-Il_
ear	.0015	7	.6739	fearsome	'f7-s@m
ear		3	.2874	earthquake	'3T-kw1k
ear		#	.0387	heart	'h#t
eau	.0001	ju	.8147	beautiful	'bju-t@-fUl
eau		5	.1853	bureau	'bj9-r5
eaux	.0000	5	1.0000	plateaux	'pl{-t5
ed	.0051	d	.6224	turned	't3nd
ed		t	.3776	looked	'lUkt
ee	.0037	i	.8728	need	'nid
ee		I	.0920	committee	k@-'mI-tI
ee		7	.0322	steering	'st7-rIN
ee		1	.0030	fiancee	fl-'qN-s1
ee#e	.0007	i	.9919	cheese	'Jiz
ee#e		7	.0081	eerie	'7-rI
eer	.0001	7	1.0000	cheers	'J7z
ei	.0017	8	.4640	heiress	'8-rIs
ei		i	.2153	receiver	rI-'si-v@R
ei		I	.2071	ageing	'1-_IN
ei		2	.0737	stein	'st2n
ei		@	.0240	foreigners	'fQ-r@-n@z
ei		1	.0119	reign	'r1n
ei		7	.0027	simultaneity	'sI-m@l-t@-'n7-tI
ei		E	.0014	leisure	'lE-Z@R
ei#e	.0000	i	.9231	receive	rI-'siv
ei#e		1	.0769	beige	'b1Z
eight	.0001	1	.8856	eight	'1t
eight		2	.1144	height	'h2t
eir	.0008	8	1.0000	their	'D8
el	.0003	P	1.0000	level	'lE-vP
en	.0016	H	1.0000	even	'i-vH
eo	.0006	i	.7017	people	'pi-pP
eo		7	.1111	theories	'T7-rIz
eo		I	.0948	pigeon	'pI-_In
eo		@	.0349	dungeon	'dVn-@n
eo		5	.0259	yeoman	'j5-m@n
eo		E	.0181	leopard	'lE-p@d
eo		Q	.0135	thereof	'D8-'rQv
eou	.0000	@	.8576	courageous	k@-'r1-@s
eou		7	.1299	hideous	'hI-d7s
eou		6	.0125	lineout	'12n-6t

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
er	.0171	@	.6253	clusters	'klV-st@z
er		3	.2166	alternative	\$l-'t3-n@-tIv
er		8	.0844	concerto	k@n-'J8-t5
er		7	.0583	adhered	@d-'h7d
er		1	.0131	dossier	'dQ-sI-1
er		#	.0023	clerk	'kl#k
er#e	.0021	8	.9000	there	'D8R
er#e		3	.1000	adverse	'{d-v3s
ern	.0001	H	1.0000	govern	'gV-vH
et	.0000	1	1.0000	beret	'bE-r1
eu	.0000	@	.5222	amateur	'{-m@-t@R
eu		7	.3011	museum	mju-'z7m
eu		3	.1040	amateurish	'{-m@-'t3-rIS
eu		4	.0727	Freudian	'fr4-d7n
eu#e	.0000	ju	.9684	queue	'kju
eu#e		3	.0316	masseuse	m{-'s3z
ew	.0009	ju	.8240	few	'fju
ew		u	.1540	blew	'blu
ew		5	.0082	sewing	's5-IN
ew		j	.0077	stewardess	'stj9-'dEs
ew		jU	.0061	steward	'stjU-@d
ew#e	.0000	j	1.0000	ewe	'ju
ey	.0018	1	.7793	surveys	's3-v1z
ey		I	.1391	money	'mV-nI
ey		2	.0677	geyser	'g2-z@R
ey		i	.0139	keyhole	'ki-h5l
ey#e	.0000	2	.9814	eye	'2
ey#e		I	.0186	honeybee	'hV-nI-bi
eye	.0002	2	1.0000	eye	'2
f	.0248	f	.6772	from	'frQm
f		v	.3228	of	'Qv
ff	.0012	f	1.0000	effect	I-'fEkt
ft	.0001	f	1.0000	often	'Q-fH
g	.0099	g	.7146	again	@-'gEn
g		_	.2821	large	'l#_
g		Z	.0033	prestige	prE-'stiZ
gg	.0003	g	.6987	aggressive	@-'grE-sIv
gg		_	.3013	suggest	s@-'_Es-t
gh	.0003	f	.9180	cough	'kQf
gh		g	.0820	ghost	'g5st
gi	.0006	_	1.0000	regionally	'ri-_H-@-II
gm	.0000	m	1.0000	paradigm	'p{-r@-d2m
gn	.0002	n	1.0000	campaign	k{m-'p1n

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
gue	.0001	g	1.0000	league	'lig
h	.0081	h	.9784	house	'h6s
h		NP	.0216	hour	'6-@R
i	.0644	I	.7733	still	'stIl
i		2	.1749	child	'J2ld
i		@	.0421	possible	'pQ-s@-bP
i		j	.0049	million	'mI-lj@n
i		i	.0048	litres	'li-t@z
i#e	.0075	2	.4687	alive	@-'l2v
i#e		I	.4618	bridge	'brI_
i#e		@	.0460	Cheshire	'JE-S@R
i#e		i	.0235	police	p@-'lis
ia	.0008	@	.6943	Belgian	'bEl-@n
ia		7	.3057	historian	hI-'st\$-r7n
ia#e	.0001	I	.6041	carriage	'k{-rI_
ia#e		7	.3959	appropriate	@-'pr5-pr7t
ie	.0022	I	.4594	babies	'b1-bIz
ie		i	.1678	field	'fild
ie		2	.1556	lies	'l2z
ie		7	.0910	Soviet	's5-v7t
ie		@	.0684	gaiety	'g1-@-tI
ie		E	.0579	friend	'frEnd
ie#e	.0004	i	.3875	believe	bI-'liv
ie#e		7	.3147	inexperience	'In-Ik-'sp7-r7ns
ie#e		I	.2313	sieve	'sIv
ie#e		@	.0664	conscience	'kQn-S@ns
ieu	.0000	Ef	.4799	lieutenant	IEf-'tE-n@nt
ieu		j3	.4503	Monsieur	m@-'sj3R
ieu		ju	.0698	lieu	'lju
iew	.0001	ju	1.0000	review	rI-'vju
igh	.0013	2	1.0000	light	'l2t
il	.0001	P	1.0000	civil	'sI-vP
in	.0003	H	1.0000	basin	'b1-sH
io		7	.5715	idiot	'I-d7t
io	.0002	@	.4286	passionate	'p{-S@-n@t
iou	.0003	7	1.0000	glorious	'gl\$-r7s
ir	.0009	3	1.0000	bird	'b3d
iu	.0000	7	1.0000	equilibrium	'i-kwI-'II-br7m
j	.0004	_	1.0000	just	'_Vst
k	.0061	k	1.0000	like	'l2k
kh	.0000	k	1.0000	khan	'k#n
kn	.0008	n	1.0000	know	'n5

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
l	.0267	l	1.0000	old	‘5ld
ld	.0017	d	1.0000	should	‘Sud
le	.0026	P	1.0000	simple	‘sIm-pP
lf	.0001	f	1.0000	half	‘h#f
li	.0001	P	1.0000	settling	‘sE-tP-IN
ll	.0059	l	1.0000	college	‘kQ-II_
lm	.0000	m	1.0000	calm	‘k#m
lv	.0000	v	1.0000	calves	‘k#vz
ly	.0000	P	1.0000	facially	‘f1-SP
m	.0256	m	1.0000	more	‘m\$R
m		F	.0000	autism	‘\$-tI-zF
mb	.0001	m	1.0000	climbed	‘kl2md
mm	.0007	m	1.0000	committee	k@-’mI-tI
mn	.0001	m	.9969	damned	‘d{md
mn		n	.0031	mnemonics	ni-’mQ-nIks
n	.0711	n	.8479	not	‘nQt
n		N	.1446	drink	‘drINk
n		H	.0075	recent	‘ri-sHt
nd	.0001	n	1.0000	grandparents	‘gr{n-’p8-r@nts
ng	.0092	N	1.0000	thing	‘TIN
ngue	.0000	N	1.0000	tongue	‘tVN
nn	.0006	n	1.0000	innocence	‘I-n@-s@ns
o	.0615	Q	.3631	body	‘bQ-dI
o		u	.1622	into	‘In-tu
o		5	.1438	most	‘m5st
o		@	.1195	London	‘lVN-d@n
o		\$	.1139	story	‘st\$-rI
o		V	.0804	covers	‘kV-v@z
o		I	.0138	women	‘wI-mIn
o		U	.0019	woman	‘wU-m@n
o		6	.0015	hour	‘6-@R
o#e	.0065	V	.2250	above	@-’bVv
o#e		5	.2226	close	‘kl5s
o#e		\$	.2000	store	‘st\$R
o#e		wV	.1773	one	‘wVn
o#e		@	.1469	purpose	‘p3-p@s
o#e		u	.0282	move	‘muv
oa	.0005	5	.7592	boat	‘b5t
oa		\$	.2408	broad	‘br\$d
oa#e	.0000	5	.5325	loathe	‘l5D
oa#e		\$	.4675	coarse	‘k\$S

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
oe	.0003	V	.5974	does	'dVz
oe		5	.2707	goes	'g5z
oe		u	.1199	shoes	'Suz
oe		i	.0097	amoeba	@-'mi-b@
oe		\$	.0024	Boer	'b\$R
oe		9	.0000	wrongdoers	'rQN-'d9z
oea	.0000	7	1.0000	diarrhoea	'd2-@-'r7
oh	.0002	5	.6489	ohms	'5mz
oh		Q	.3511	John	'_Qn
oi	.0005	4	.9803	avoid	@-'v4d
oi		w#	.0197	reservoir	'rE-z@-vw#R
oi#e	.0001	4	.9673	choice	'J4s
oi#e		w#	.0327	armoire	'#m-w#R
ol	.0000	P	1.0000	pistol	'pI-stP
on	.0010	H	1.0000	person	'p3-sH
oo	.0028	U	.4660	good	'gUd
oo		u	.4175	food	'fud
oo		\$	.0874	door	'd\$R
oo		V	.0254	blood	'bIVd
oo		5	.0036	brooch	'br5J
oo		9	.0001	boorish	'b9-rIS
oo#e	.0000	u	.8101	choose	'Juz
oo#e		U	.1899	goodie	'gU-dI
or	.0112	\$	.5365	former	'f\$-m@R
or		R	.2562	mayor	'm8R
or		@	.1272	effort	'E-f@t
or		3	.0801	work	'w3k
or#e	.0001	\$	1.0000	force	'f\$S
ou	.0101	6	.3558	around	@-'r6nd
ou		u	.2173	group	'gruP
ou		U	.1697	should	'SUd
ou		@	.1004	conscious	'kQn-S@s
ou		\$	.0721	four	'f\$R
ou		V	.0643	country	'kVn-trI
ou		9	.0087	tenuous	'tE-nj9s
ou		5	.0065	soul	's5I
ou		3	.0044	journal	'_3-nP
ou		Q	.0008	cough	'kQf
ou#e	.0007	6	.4074	house	'h6s
ou#e		\$	.3197	course	'k\$S
ou#e		9	.1538	bourse	'b9s
ou#e		u	.1181	route	'rut
ou#e		3	.0009	scourge	'sk3_
ough	.0010	\$	.4262	thought	'T\$t
ough		u	.2758	through	'Tru
ough		5	.2455	though	'D5
ough		6	.0353	ploughing	'pl6-IN
ough		@	.0171	borough	'bV-r@

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
our	.0006	\$	.6729	court	'k\$t
our		@	.2516	colours	'kV-l@z
our		3	.0684	journey	'_3-nI
our		9	.0071	tourney	't9-nI
oux	.0000	u	1.0000	Sioux	'su
ow	.0040	6	.4422	down	'd6n
ow		5	.4320	know	'n5
ow		@	.1017	arrowroot	'{-r@-rut
ow		Q	.0241	knowledge	'nQ-II_
ow#e	.0000	5	.6286	owe	'5
ow#e		6	.3715	browse	'br6z
oy	.0003	4	1.0000	joy	'_4
oy#e	.0000	4	1.0000	Joyce	'_4s
p	.0180	p	1.0000	people	'pi-pP
pb	.0000	b	1.0000	cupboard	'kV-b@d
ph	.0004	f	.9878	telephone	'tE-II-f5n
ph		p	.0068	shepherds	'SE-p@dz
ph		v	.0054	nephew	'nE-vju
pn	.0000	n	1.0000	pneumonia	nju-'m5-nj@
pp	.0012	p	1.0000	apple	'{-pP
pph	.0000	f	1.0000	sapphire	's{-f2-@R
ps	.0000	s	1.0000	psychological	's2-k@-'lQ-_I-kP
pt	.0000	t	1.0000	receipt	rI-'sit
q	.0010	k	1.0000	quite	'kw2t
qu	.0001	k	1.0000	technique	tEk-'nik
r	.0482	r	.5808	real	'r7l
r		R	.4192	car	'k#R
re	.0059	@R	1.0000	centre	'sEn-t@R
rh	.0000	r	1.0000	rhythm	'rI-D@m
rr	.0009	r	.9962	terrible	'tE-r@-bP
rr		R	.0038	bizarre	bl-'z#R
s	.0607	s	.5615	last	'l#st
s		z	.4253	these	'Diz
s		Z	.0076	pleasure	'plE-Z@R
s		S	.0055	sure	'S\$R
sc	.0001	s	.9807	scientific	's2-@n-'tI-flk
sc		S	.0108	fascism	'f{-SI-z@m
sc		k	.0085	viscount	'v2-k6nt
sch	.0002	s	.9587	schism	'sI-z@m
sch		S	.0413	schedule	'SE-djul

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
sh	.0036	S	1.0000	fish	'fIS
si	.0005	Z	.5457	vision	'vI-ZH
si		S	.4543	version	'v3-SH
sl	.0000	l	1.0000	island	'2-l@nd
ss	.0026	s	.8887	possible	'pQ-s@-bP
ss		S	.0674	issues	'I-Suz
ss		z	.0439	possession	p@-'zE-SH
st	.0001	s	1.0000	castle	'k#-sP
sw	.0001	s	1.0000	answer	'#n-s@R
t	.0686	t	.9405	time	't2m
t		S	.0442	initiate	I-'nI-SI-It
t		J	.0153	picture	'pIk-J@R
tch	.0003	J	1.0000	dispatch	dI-'sp{J
tg	.0000	g	1.0000	mortgage	'm\$-gI_
th	.0345	D	.8832	other	'V-D@R
th		T	.1097	nothing	'nV-TIN
th		t	.0070	thirdly	't3d-II
ti	.0031	S	.9441	emotions	I-'m5-SHz
ti		J	.0536	question	'kwEs-J@n
ti		Z	.0023	equation	I-'kw1-ZH
tsch	.0000	J	1.0000	putsch	'pUJ
tt	.0018	t	1.0000	attention	@-'tEn-SH
tw	.0004	t	1.0000	two	'tu
u	.0184	V	.4135	but	'bVt
u		ju	.2063	united	ju-'n2-tId
u		jU	.1298	argument	'#-gjU-m@nt
u		@	.0909	until	@n-'tIl
u		w	.0562	language	'l{N-gwI_
u		U	.0530	full	'fUl
u		u	.0295	revolution	'rE-v@-'lu-SH
u		I	.0198	busy	'bI-zI
u		E	.0011	burial	'bE-r7I
u#e	.0021	ju	.4202	abuse	@-'bjus
u#e		@	.2435	future	'fju-J@R
u#e		V	.1262	bulge	'bVl_
u#e		u	.0820	absolute	'{b-s@-lut
u#e		3	.0573	nurse	'n3s
u#e		\$	.0554	assure	@-'S\$R
u#e		I	.0144	minute	'mI-nIt
u#e		9	.0009	lure	'l9R
ua	.0004	9	1.0000	actual	'{k-J9I
uar	.0000	#	1.0000	guard	'g#d
ue	.0004	j	.5901	values	'v{l-juz
ue		u	.4099	issue	'I-Su

APPENDIX (Continued)

Grapheme	PP	Phoneme	Prob	Word	Pronunciation
ui	.0006	I	.7331	build	'bIId
ui		u	.1948	juicy	'_u-sI
ui		ju	.0721	pursuit	p@-'sjut
ui#e	.0000	u	1.0000	juice	'_us
ul	.0000	P	1.0000	faculty	'f{-kP-tI
uo	.0000	4	1.0000	buoyant	'b4-@nt
ur	.0009	3	.6860	absurd	@b-'s3d
ur		@	.2898	Saturday	's{-t@-dI
ur		9	.0242	lured	'I9d
ur#e	.0001	3	1.0000	curve	'k3v
uy	.0001	2	1.0000	buy	'b2
v	.0112	v	1.0000	view	'vju
w	.0143	w	1.0000	week	'wik
wh	.0044	w	.7963	which	'wIJ
wh		h	.2037	who	'hu
wr	.0003	r	1.0000	wrong	'rQN
x	.0019	ks	.8228	experience	Ik-'sp7-r7ns
x		gz	.1646	example	Ig-'z#m-pP
x		kS	.0126	sexual	'sEk-S9I
xh	.0000	gz	1.0000	exhaustion	Ig-'z\$s-J@n
y	.0170	I	.6181	mysterious	mI-'st7-r7s
y		j	.2156	beyond	bI-'jQnd
y		2	.1592	try	'tr2
y		@	.0067	analysis	@-'n{-l@-sIs
y		i	.0004	lycee	'li-sI
y#e	.0002	I	.6432	tricycle	'tr2-sI-kP
y#e		2	.3568	type	't2p
ya	.0000	7	1.0000	Aryan	'8-r7n
z	.0007	z	.9676	size	's2z
z		s	.0305	blitz	'bIIts
z		Z	.0019	seizure	'si-Z@R
zz	.0000	z	.9571	puzzled	'pV-zPPd
zz		ts	.0429	mezzanine	'mE-ts@-nin

(Manuscript received March 27, 2000;
revision accepted for publication October 2, 2001.)