TRANSPARENCY MEASURES FOR BRITISH ENGLISH

Feed-Forward, -Backward and Neutral Transparency Measures for British English

Kenneth A. Spencer

Institute for Learning, University of Hull, U.K.

Correspondence concerning this article should be addressed to K. A. Spencer, Institute for Learning, University of Hull, HU6 7RX, U.K. (e-mail: k.a.spencer@hull.ac.uk).

Abstract

Orthographic transparency metrics for opaque or deep languages such as French and English have tended to focus on feedforward and/or feedback directions, with claims made for the influence of both on reading. In the present study, data for five transparency metrics for Southern British English, three of which are neither feedforward nor feedback, are presented demonstrating the complex relationships between metrics, and offering an explanation for feedback effects in children's reading accuracy. The structure of such metrics from a variety of corpus sizes and origins is investigated, concluding that large corpus sizes do not make a substantial contribution to the value of such metrics when compared with smaller samples, and that adult and child corpuses have very similar profiles.

The regularity of the written forms of languages varies, and those orthographies (Scheerer, 1986) that are regular, usually with one-to-one mapping of phonemes and letters, are classed as “transparent” or “shallow”, and those with irregular correspondences are classed as “opaque” or “deep”[1]. Finnish and Turkish are highly transparent for both reading and spelling, with one-to-one mapping of graphemes and phonemes. Greek and German are less transparent for spelling than reading, with letters that have clearly defined pronunciations, but phonemes that have alternative spellings. English and French are opaque for both (Goswami, Porpodas, & Wheelwright, 1997). It has been proposed (Frost, Katz, & Bentin, 1987; Katz & Frost, 1992) that the variation between orthographies leads to processing differences for naming and lexical decision, and literacy acquisition. Transparent orthographies are certainly easier to learn (Cossu, Shankweiler, Liberman, Katz, & Tola, 1988; Oney & Goldman, 1984; Seymour, Aro and Erskine, 2003; Spencer & Hanley, 2003), and comparisons for adults between opaque (English; French) and transparent (Italian) languages demonstrate a clear processing advantage for Italian (Paulesu et al, 2000), and especially for dyslexic adults (Paulesu et al, 2001). For normal and dyslexic children, accuracy levels are lower and reading speed slower for deeper languages (Cossu, Gugliotta, & Marshall, 1995; Ellis et al., 2004; Frith, Wimmer, & Landerl, 1998; Landerl, Wimmer, & Frith, 1997). Seymour et al. (2003) found an abrupt rather than a graded effect of transparency for the 13 orthographies included in their study, and they concluded that there is a threshold of orthographic transparency which, once exceeded, results in a step change in literacy acquisition. However, this may be because transparency variations have not always been subjected to comprehensive linguistic analyses that allow languages to be placed along a finely graded continuum. Seymour et al. simply used a hypothetical classification of

European orthographies based on the team’s estimates of the extent of variability in their languages.

More detailed linguistic analyses have measured the orthographic body/phonological rime transparency of French (Ziegler, Jacobs, & Stone, 1996), and English (Ziegler, Stone, & Jacobs, 1997) for both reading and spelling, and allow comparisons of relative transparency. When compared with English, French is 20% more consistent for reading, but 10% less so for spelling.

This large-grain, body/rime level of analysis is central to the connectionist or parallel distributed processing network model of reading (Plaut, McClelland, Seidenberg, & Patterson, 1996). In this model, the ease with which a word is pronounced depends on its relative orthographic body transparency. Consistent words have body letter patterns that have the same phonological rime and are always pronounced in the same way. For inconsistent words, the less typical the pronunciation, the greater the word reading difficulty. In this scheme, a simple, dichotomous classification may be applied to the transparency of individual words (consistent/inconsistent), but a more sophisticated categorisation that refers to regularity and consistency (see Jared, 2002) may also be appropriate. Treiman, Mullennix, Bijeljac-Babic, and Richmond-Welty (1995) see this as inadequate, and claim that only continuous representations of consistency allow detailed examination of transparency effects. This has led to denials of a strict dichotomy between words in the connectionist approach, with Plaut (1999) suggesting that “language knowledge is inherently graded, and the language mechanism is a learning device that gradually picks up on statistical structure among written and spoken words and the contexts in which they occur” (p. 544).

The application of body/rime reading and spelling transparency probabilities for English and French (Ziegler et al., 1996, 1997) have led to a controversial interpretation of the effects of transparency. For many years it has been assumed that visual word perception is influenced in a feedforward direction. In other words, reading transparency influences reading, and spelling transparency influences spelling. However, Stone, Vanhoy and Van Orden (1997) challenged this assumption, claiming that both spelling (feedback) and reading (feedforward) transparency influences reading in English. Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) see feedback rime effects as supporting a highly interactive system in which both spelling and reading patterns contribute to the naming process as it unfolds across time. Norris, McQueen and Cutler (2000) disagree, suggesting that this effect is caused by feedforward models being sensitive to the tendency of inconsistent body-rimes to have lower type frequencies. Kessler, Treiman and Mullennix (2007) have reviewed the evidence for this feedback effect and remain sceptical, concluding that the results are surprising because "inconsistency in the sound-to-letter direction, something that might logically make writing difficult, would seem to play no necessary role in reading, which involves mapping letters to sounds" (p. 159).

Although recent studies have focussed on body/rime transparency, English has also been extensively studied at the fine-grain, grapheme-phoneme level for reading and spelling (Berndt, Reggia, & Mitchum, 1987; Carney, 1994; Gontijo, Gontijo, & Shillcock, 2003; Hanna, Hanna, & Hodges, 1966; Venezky, 1967). Venezky (1970) introduced a series of rules, based on his analysis of 20,000 words, for the pronunciation of English orthography, but noted that there was little known about the processes involved in the learning of these major patterns. Although he eschewed the idea of probability tables for his seminal work, he did warn that more types of relationships than the simple regular-irregular were needed to adequately describe English orthography.

However, it was the regular-irregular dichotomy that was investigated by Baron and Strawson (1976) who found that regular words that conformed to Venezky's rules are read aloud more quickly than words that do not conform (exceptions). There have since been numerous studies demonstrating longer naming for exception words (Hino & Lupker, 2000; Stanovich & Bauer, 1978; Waters & Seidenberg, 1985). Although grapheme-phoneme correspondence probabilities have been used to describe the transparency of word components (Reggia, Marsland, & Berndt, 1988), it is dichotomous grapheme-phoneme correspondence rules that represent spelling-sound knowledge in one of the main theoretical approaches to the development of computer-based models of reading (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). Words are divided into those that obey the rules, the regular (transparent) words, and those that are irregular (opaque), the two categories being processed in different ways - a view similar to that expressed by Seymour et al. (2003) for comparisons between languages.

Hanna et al. (1966) also worked with a large corpus of words (17,310) but their emphasis was on spelling and the specification of phoneme-grapheme correspondences. Unlike Venezky, their results were presented as proportions from which spelling conditional probabilities could be calculated. Berndt et al. (1987) later re-worked the Hanna et al. data to produce a similar set of grapheme-phoneme correspondence (reading) probabilities that formed the basis for the dual-route connectionist model of reading (Reggia, et al., 1988). More recently Perry, Ziegler and Coltheart (2002) have referred to such probabilities as contingency measures.

Variations of fine-grain word transparency within a deep language such as English provide insights into literacy acquisition processes, and because of this they formed an integral part of the early serial dual route models of skilled reading (Coltheart, Curtis, Atkins, & Haller, 1993). The computer-based studies divided words into graphemes and their associated phonemes, which were termed grapheme-phoneme correspondences (GPC) when associated with studies of reading, and phoneme-grapheme correspondences (PGC) when associated with spelling. Separate GPC and PGC conditional probabilities are calculated from the same sound-letter components. For example, the PGCs derived from Hanna et al.'s (1966) study, and used to calculate PGC probabilities, were used by Berndt et al. (1987) to calculate GPC conditional probabilities.

This approach to the generation of reading correspondence probabilities is evident in early versions of the serial dual route model of reading (Coltheart et al., 1993) that derived GPC values from a learning algorithm. The hope was that this would mimic the human processes for developing GPC rules used in the interpretation of new words. However, the results were subject to a distortion that had been noted by Berndt et al. (1987), who demonstrated that within their conditional probability reading metric, infrequent graphemes often had very high reading probabilities. For example, "the word calf is highly predictable (*m* = 1.0) . . . . Yet because the grapheme LF occurs infrequently (p = .00008), the correspondence LF → f, though a consistent pronunciation for that grapheme, may not be easily learned by beginning readers or well retained by adults with reading disorders" (p. 4). In effect, obscure graphemes that were always pronounced the same way dominated the processing procedures of the model because of their high conditional probabilities, and in order to exclude such effects Coltheart et al. took the expedient of applying a minimum frequency rule to their algorithm. Eventually it was conceded that algorithm-derived rules did not work for the DRC model and that there was a lack of knowledge concerning the development in humans of the DRC model's GPC rules, which are but a 'set of hypotheses about what GPC rules skilled readers possess' (Rastle & Coltheart, 1999, p. 484). Perry, Ziegler and Zorzi (2007) claim that this absence of learning within the model means that it cannot be used to simulate reading development and developmental reading disorders, and has prompted the development of their CDP+

model's delta rule learning that has psychological reality in the form of a classical conditioning law.

The corpus size and the source of the material (e.g. adult or child) may influence the outcome when calculating word metrics. Venezky (1967) and Hanna et al. (1966) used very substantial multisyllabic adult corpuses of 20,00 and 17,310 words, whereas Ziegler et al. (1996, 1997) used smaller samples of monosyllabic words (1,843; 2,694), as did Coltheart et al. (1993), basing their calculations on the 2,897 monosyllabic corpus of Seidenberg and McClelland (1989). McGuinness (1998) criticised both Hanna et al. and Venezky for using large corpuses when determining the probability structure of English, suggesting an optimum corpus size of 3,000 words because spelling patterns for smaller corpuses of more common words would differ from more extensive analyses and have greater utility. This paper considers the impact of both corpus size and source material on the calculation of a multisyllabic word metrics.

The present paper also suggests that for foundation literacy, small-grain analyses of English words, at the phoneme level, provide powerful predictors of single word decoding and may provide insights into the acquisition of a knowledge base that informs early reading processes. Counter-intuitively, it proposes that single word decoding, for children, is influenced by feedback (spelling) transparency, rather than feedforward (reading) transparency (see Spencer, 2001) because there is a more fundamental association that is independent of the direction of the transparency. However, this depends on the way in which transparency metrics are derived, which in turn necessitates further consideration of the way in which reading and spelling transparency and probability/contingency are calculated.

Grain Consistency, Correspondences and Probabilities

As we have seen, in Berndt et al.'s (1987) study, graphemes were mapped on to phonemes to produce GPC conditional probabilities (G → P); these probabilities were derived from Hanna et al.'s original data that had calculated PGC conditional probabilities (P → G). However, both probabilities are derived from an underlying association that seldom features in models of reading or spelling, but is the small grain equivalent of the metric that gives rise to the type frequency (Norris et al., 2000) that explains feedback effects. Before conditional probabilities are calculated, the data on which they are based exist as a directionless association between graphemes and phonemes, each of which has a type frequency in the corpus from which it is derived (P ↔ G or G ↔ P). It is this type frequency that is used to calculate either its GPC or PGC probability value. In this paper the term *sonograph* is used to describe this directionless item.

An example: calculating 5 metrics for /eɪ/ ↔ <eigh> from Hanna et al. (1966)

The values for the example sonograph are summarised in Table 1.

TABLE 1 about here

Phoneme and Grapheme frequency and probability of occurrence. It is obvious that graphemes occur at different rates in an opaque language, but perhaps less obvious that phonemes also have considerable variation in frequency. For the Hanna corpus, adapted for the simplified phoneme structure proposed by Fry (2004), the grapheme type frequency range is from 9,119 to 1, and the phoneme range from 9,390 to 102. In this example, the grapheme <eigh> has a type frequency of 51, and the phoneme /eɪ/ a frequency of 2,248. Because there is a total of 108,571 sonographs in the Hanna corpus of 17,310 words, the probability of occurrence is 0.0005 (51/108571) for the grapheme <eigh>, and 0.0207 (2248/108571) for the phoneme /eɪ/.

Sonograph probability of occurrence. The PGC /eɪ/ → <eigh>, as in *eight*, occurs 48 times in the Hanna et al. (1966) corpus, as does the GPC <eigh> → /eɪ/. Thus, we can simply say that /eɪ/ ↔ <eigh> has a type frequency of 48. Therefore, the sonograph /eɪ/ ↔ <eigh> has a probability of occurrence of 0.0004 (48/108571) within the corpus. Each of the 362 unique sonograph mappings also has an associated frequency in the corpus, and it is from these frequencies that separate spelling and reading probabilities are calculated.

GP correspondence probability. The grapheme <eigh> appears 51 times in the corpus and is associated with two phonemes, /aɪ/ ↔ <eigh> (as in *height;* type frequency = 3) and /eɪ/ ↔ <eigh> (as in *eight;* type frequency = 48). Therefore, sonograph /eɪ/ ↔ <eigh> has a conditional reading probability of 0.9412 (48/51).

Phoneme-grapheme (spelling) probability. The phoneme /eɪ/ appears 2,248 times in the corpus and is associated with 14 graphemes. Within this group of representations of the single phoneme, the sonograph /eɪ/ ↔ <eigh> has a frequency of 48 and a PGC spelling probability of 0.0214 (48/2248).

METHOD

Corpuses

For the present analysis a database was generated from the most frequent 7,000 words in the Lancaster-Oslo-Bergen (Hofland & Johansson, 1982) adult word count. The words were partially lemmatised by removing inflected forms (such as: cleaned, cleaning, cleaned) if the base word was included in the list; other inflected forms were retained. This process resulted in a reduced sample of 3,220 multisyllabic words. In order to study the changing nature of the metrics for increasingly larger corpuses, this group was divided into two smaller groups: 928 words with a frequency per million greater then 80; 2,019 words with a frequency per million greater than 30.

A second database was generated from the most frequent 1,000 words in the Children's Printed Word Database (Masterson, Stuart, Dixon & Lovejoy, 2003), which was reduced to 971 multisyllabic words after partial lemmatising and removing abbreviations. Phonology codes were based on the on-line version of the Oxford English Dictionary (Oxford University Press, 2006), which provides a balance between broad and narrow transcriptions.

Word Decomposition

All words were scanned individually by a computer programme designed to identify likely orthographic representations for phonemes, based on forms identified by Spencer (1999). The final version of alignments for all words was visually inspected, and corrected for minor inconsistencies. This method is slightly different to that used by Gontijo et al. (2003). For example, the grapheme /a-e/ is only used for words in which the distance between the *a* and *e* is one letter (*face*); for greater distances the *e* is associated with its contiguous letters. This method has been shown to have greater predictive power in regression analyses of spelling and reading than the Gontijo et al. method (Spencer, 2007, 2008).

The alignment process resulted in a total of 17,058 grapheme-phoneme pairs (sonographs) for the adult corpus and 3,814 pairs for the children's corpus. The pairs were entered into Excel spreadsheets that were programmed to provide counts of the five word metrics outlined above. Both the adult and children's corpuses included all 44 phonemes within the Oxford English Dictionary transcriptions, which represent a southern British phonology[2]. For the full adult corpus 168 graphemes were identified, and 122 for the children's corpus, producing 316 unique adult sonograph mappings and 217 children's mappings. Frequency counts for each mapping were produced, from which reading and spelling probabilities were calculated.

RESULTS AND DISCUSSION

The metrics for each corpus are presented in Appendices A to C. Table 2 shows that the values for the five metrics derived from the four sources are highly inter-correlated. The mean correlations among the sources are: .96 for grapheme and phoneme probabilities of occurrence, and grapheme-phoneme (reading) correspondence probabilities; .93 for phoneme-grapheme (spelling) correspondence probabilities; and .88 for sonograph probabilities of occurrence. This suggests that the metrics describing British English remain substantially the same for corpuses of varying size, from sources of writing for adults or children. The present analysis finds that the core probability values for English may be obtained from relatively small samples of words, and this is reflected in Figure 1. The children's 1K corpus of 217 sonographs has 139 (64%) in the relative probability[3] of occurrence range .01 to 1.0 (shown between A and B, Figure 1), with the remaining 78 (36%) sonographs having a probability of occurrence that is less than .01. The adult 1K corpus has 18 additional sonographs, all of which are of low probability (< 0.01), indicated by the bars between B and C, with 45% < .01. Thus, the profiles of the adult and children's corpuses are very similar, with the addition of a small number of very infrequent sonographs in the adult corpus. This pattern is repeated for the increasingly larger adult corpuses. The difference between adult corpus size of 1K and 2K is an additional 32 very low frequency sonographs (the bars between B and C), with 51% < .01; and between 2K and 3K, an additional 49 low frequency sonographs, with 59% < .01. For larger corpuses the long tail of very low frequencies increases, for Hanna et al. (1966) 74% of 362 sonographs have relative frequencies < .01; and for Gontijo et al. 69% out of 461.

The purpose of calculating descriptive metrics for an opaque language such as English is to provided continuous measures of the variables that may have relevance for linguistic and psychological studies. This has usually involved the decomposition of very large corpuses, but it appears that McGuinness (1998) is correct in assuming that relatively small corpuses can also yield metrics of great utility. Spencer (2008) demonstrated that smaller samples of words produced metrics that predicted a larger proportion of the variance in young children's reading than those from larger samples, and that for reading, sonograph probability predicted more of the variance than spelling probability, whereas reading probability made no contribution to the variance.

Table 2 also demonstrates several relationships between variables that may account for the failure of reading metrics to operate as continuous measures in serial processing computer models, and for spelling metrics to have a feedback influence in reading. The substantial negative correlation between grapheme probability of occurrence and grapheme-phoneme correspondence probability reflects the unfortunate nature of the reading feedforward metric: more frequent graphemes tend to have values less than 1.0 because they are associated with more than one phoneme, but infrequent graphemes tend to have probabilities close to 1.0 because they often are associated with only one phoneme. It is this aspect of the metric that was noticed by Berndt et al. (1987) and commented upon by Coltheart et al. (1993). There is a similar but smaller association between phoneme and phoneme-grapheme correspondence probabilities. It is smaller because there are fewer phonemes than graphemes, and there is less variation in their frequency. In contrast to this, there is a very significant, positive association between sonograph probability of occurrence and phoneme-grapheme correspondence probability. This reflects the fact that high frequency sonographs tend to have high phoneme-grapheme probabilities. The highly significant association between sonograph frequency and spelling probabilities may account for the apparent feedback effects that have been observed, as Norris et al. (2000) proposed.

TABLE 2 about here

FIGURE 1 about here

CONCLUSION

Since the time of early computer analyses of English (Venezky, 1967; Hanna et al., 1966) word metrics have tended to focus on the proportional values of phonemes per grapheme (reading) or graphemes per phoneme (spelling). This approach has been extended to produce a plethora of terms associated with different grain and corpus sizes (e.g. regularity, consistency, neighbourhood, friends, enemies, conditional probability, contingency). In the present paper the primary fine-grain metrics, from which many secondary metrics are calculated, are included to provide researchers, especially in the field of children's psycholinguistics, with a range of measures that may be incorporated into models of reading and spelling to gain a more comprehensive understanding of the development of foundation literacy skills in an opaque language.

The results suggest that metrics derived from adult corpuses, that are incorporated into computational models, are appropriate for use within children's developmental and learning studies, but that inappropriate methods of metric calculation, e.g. GPC values, may have inhibited the development of learning algorithms within such models. The association between spelling (feedback) probabilities and sonograph probabilities may offer an explanation of the contentious feedback effects that have been observed in reading studies. It may well be the case that children's knowledge of correspondences is simply a frequency effect (Spencer, 2008), and that acquisition of such knowledge is subject to the simple conditioning processes advocated by Perry et al (2007) in their CDP+ model.

APPENDIX A: Phoneme probability of occurrence

Corpus A is derived from Hofland, K. & Johansson, S. (1982): 3K = 3,220 words; 2K = 2,019 words; 1K = 928 words.

Corpus C is derived from Masterson, J., Stuart, M., Dixon, M. & Lovejoy, S. (2003): 1K = 971 words.

			Corpus A			Corpus C
Phoneme	Example Grapheme	Example Word	3K	2K	1K	1K
æ	a	and	0.0188	0.0175	0.0164	0.0218
ɑː	are	are	0.0075	0.0084	0.0093	0.0100
b	b	be	0.0179	0.0173	0.0179	0.0281
k	c	can	0.0476	0.0472	0.0409	0.0467
tʃ	ch	which	0.0067	0.0074	0.0082	0.0073
d	d	and	0.0363	0.0367	0.0384	0.0456
iː	e	he	0.0345	0.0362	0.0382	0.0317
ə	e	the	0.0856	0.0804	0.0741	0.0396
e	e	when	0.0331	0.0350	0.0332	0.0273
ɜː	er	her	0.0075	0.0069	0.0086	0.0055
ɪə	ere	here	0.0062	0.0061	0.0082	0.0034
eə	ere	there	0.0036	0.0037	0.0034	0.0050
eɪ	ey	they	0.0193	0.0193	0.0204	0.0205
f	f	for	0.0208	0.0222	0.0220	0.0233
g	g	get	0.0104	0.0107	0.0095	0.0157
h	h	he	0.0067	0.0074	0.0114	0.0163
ɪ	i	in	0.0711	0.0678	0.0582	0.0600
aɪ	i	mind	0.0146	0.0160	0.0164	0.0202
dʒ	j	just	0.0089	0.0087	0.0073	0.0076
l	l	like	0.0596	0.0591	0.0584	0.0569
m	m	from	0.0322	0.0323	0.0345	0.0302
ŋ	ng	bang	0.0049	0.0048	0.0057	0.0165
n	n	and	0.0791	0.0783	0.0766	0.0572
ɒ	o	of	0.0158	0.0150	0.0182	0.0218
əʊ	o	so	0.0109	0.0118	0.0136	0.0184
uː	o	to	0.0102	0.0107	0.0109	0.0126
ɔɪ	oi	point	0.0014	0.0015	0.0018	0.0018
ʊə	oor	poor	0.0021	0.0023	0.0020	0.0005
ɔː	or	for	0.0103	0.0111	0.0145	0.0128
aʊ	ou	about	0.0041	0.0046	0.0070	0.0089
p	p	up	0.0340	0.0342	0.0348	0.0359
r	r	from	0.0502	0.0484	0.0411	0.0380
ʒ	s	usually	0.0015	0.0013	0.0011	0.0005
s	s	this	0.0645	0.0650	0.0641	0.0582
z	s	is	0.0103	0.0096	0.0114	0.0317
ʃ	sh	she	0.0168	0.0157	0.0125	0.0094
t	t	to	0.0740	0.0748	0.0747	0.0745
ð	th	the	0.0028	0.0036	0.0073	0.0060
θ	th	think	0.0048	0.0056	0.0064	0.0063
ə	u	put	0.0037	0.0040	0.0045	0.0058
ʌ	u	but	0.0140	0.0141	0.0168	0.0225
v	v	very	0.0163	0.0163	0.0164	0.0115
w	w	was	0.0109	0.0124	0.0145	0.0215
j	y	you	0.0084	0.0086	0.0093	0.0050

APPENDIX B: Grapheme and GPC Probability

Grapheme probability = Grapheme probability of occurrence.

GPC probability = Grapheme-phoneme correspondence (reading) probability.

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
a	eɪ	waste	0.0567	0.0540	0.0522	0.0482	0.1399	0.1335	0.1304	0.1033
a	ə	woman	0.0567	0.0540	0.0522	0.0482	0.3953	0.3648	0.3304	0.1685
a	eə	vary	0.0567	0.0540	0.0522	-	0.0072	0.0071	0.0087	-
a	ɒ	whatever	0.0567	0.0540	0.0522	0.0482	0.0133	0.0160	0.0304	0.0652
a	æ	whereas	0.0567	0.0540	0.0522	0.0482	0.3309	0.3238	0.3130	0.4511
a	ɑː	vast	0.0567	0.0540	0.0522	0.0482	0.0562	0.0676	0.0739	0.0870
a	ɪ	village	0.0567	0.0540	0.0522	0.0482	0.0255	0.0356	0.0348	0.0217
a	ɔː	false	0.0567	0.0540	0.0522	0.0482	0.0235	0.0374	0.0565	0.0815
a	e	necessarily	0.0567	0.0540	0.0522	0.0482	0.0082	0.0142	0.0217	0.0217
ach	ɒ	yacht	0.0001	-	-	-	1.0000	-	-	-
ae	iː	aesthetic	0.0001	-	-	-	1.0000	-	-	-
a-e	eɪ	wave	0.0056	0.0058	0.0066	0.0073	1.0000	1.0000	1.0000	1.0000
ai	eɪ	wait	0.0035	0.0037	0.0032	0.0034	0.8525	0.8718	0.6429	0.6923
ai	ə	uncertainty	0.0035	0.0037	0.0032	0.0034	0.0820	0.0513	0.1429	0.0769
ai	eə	dairy	0.0035	-	-	0.0034	0.0164	-	-	0.0769
ai	e	said	0.0035	0.0037	0.0032	0.0034	0.0492	0.0769	0.2143	0.1538
aigh	eɪ	straight	0.0001	0.0001	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
air	eə	upstairs	0.0009	0.0009	0.0009	0.0013	1.0000	1.0000	1.0000	1.0000
al	ɔː	walk	0.0048	0.0051	0.0048	0.0016	0.0244	0.0377	0.0952	0.5000
al	l	vital	0.0048	0.0051	0.0048	0.0016	0.9756	0.9623	0.9048	0.5000
ar	ə	upwards	0.0057	0.0062	0.0073	0.0055	0.2551	0.2344	0.2813	0.0952
ar	eə	scarcely	0.0057	-	-	-	0.0102	-	-	-
ar	ɑː	yard	0.0057	0.0062	0.0073	0.0055	0.6633	0.6875	0.6563	0.8095
ar	ɔː	warmth	0.0057	0.0062	0.0073	0.0055	0.0714	0.0781	0.0625	0.0952
are	eə	welfare	0.0012	0.0013	0.0011	0.0013	0.9524	0.9286	0.8000	0.8000
are	ɑː	are	0.0012	0.0013	0.0011	0.0013	0.0476	0.0714	0.2000	0.2000
au	ə	restaurant	0.0012	0.0012	-	-	0.0500	0.0769	-	-
au	ɒ	because	0.0012	0.0012	0.0009	0.0016	0.0500	0.0769	0.2500	0.1667
au	ɑː	laughter	0.0012	0.0012	0.0009	0.0016	0.3000	0.2308	0.2500	0.6667
au	ɔː	pause	0.0012	0.0012	0.0009	0.0016	0.6000	0.6154	0.5000	0.1667
augh	ɔː	taught	0.0002	0.0003	0.0005	0.0005	1.0000	1.0000	1.0000	1.0000
aur	ɔː	dinosaur	-	-	-	0.0003	-	-	-	1.0000
aw	ɔː	withdrawn	0.0007	0.0005	0.0005	0.0008	1.0000	1.0000	1.0000	1.0000
ay	eɪ	yesterday	0.0017	0.0019	0.0032	0.0047	1.0000	1.0000	1.0000	1.0000
ayer	eə	prayer	0.0001	-	-	-	1.0000	-	-	-
ayor	eə	mayor	0.0001	-	-	-	1.0000	-	-	-

APPENDIX B: (Continued)

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
b	b	whereby	0.0177	0.0173	0.0179	0.0270	1.0000	1.0000	1.0000	1.0000
bb	b	rubber	0.0001	-	-	0.0010	1.0000	-	-	1.0000
bt	t	undoubtedly	0.0003	0.0002	0.0002	-	1.0000	1.0000	1.0000	-
c	ʃ	appreciation	0.0388	-	-	-	0.0030	-	-	-
c	k	welcome	0.0388	0.0376	0.0350	0.0270	0.8132	0.8210	0.7727	0.8155
c	s	velocity	0.0388	0.0376	0.0350	0.0270	0.1839	0.1790	0.2273	0.1845
cc	k	occurrence	0.0009	0.0005	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
ce	s	voice	0.0050	0.0053	0.0057	0.0013	1.0000	1.0000	1.0000	1.0000
ch	ʃ	machinery	0.0047	0.0053	0.0061	0.0068	0.0247	0.0182	0.0370	0.0385
ch	tʃ	which	0.0047	0.0053	0.0061	0.0068	0.7407	0.8182	0.8148	0.8077
ch	k	technology	0.0047	0.0053	0.0061	0.0068	0.2346	0.1636	0.1481	0.1538
ci	ʃ	suspicion	0.0012	0.0012	0.0011	0.0003	1.0000	1.0000	1.0000	1.0000
ck	k	wicked	0.0022	0.0026	0.0011	0.0060	1.0000	1.0000	1.0000	1.0000
cq	k	acquisition	0.0001	-	-	-	1.0000	-	-	-
d	dʒ	soldier	0.0358	0.0362	0.0379	-	0.0097	0.0133	0.0120	-
d	d	yield	0.0358	0.0362	0.0379	0.0420	0.9903	0.9867	0.9880	0.9875
d	t	chased	-	-	-	0.0420	-	-	-	0.0125
dd	d	wedding	0.0007	0.0010	0.0007	0.0005	1.0000	1.0000	1.0000	1.0000
dg	dʒ	judgment	0.0001	0.0002	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
dge	dʒ	ridge	0.0005	0.0004	0.0002	0.0008	1.0000	1.0000	1.0000	1.0000
dj	dʒ	adjustment	0.0001	-	-	-	1.0000	-	-	-
e	ə	witness	0.0649	0.0661	0.0588	0.0383	0.2161	0.2023	0.1815	0.1233
e	ɪə	zero	0.0649	0.0661	0.0588	0.0383	0.0143	0.0131	0.0232	0.0068
e	eə	wherever	0.0649	-	-	-	0.0009	-	-	-
e	ɪ	women	0.0649	0.0661	0.0588	0.0383	0.2723	0.2693	0.2432	0.2260
e	iː	we	0.0649	0.0661	0.0588	0.0383	0.0375	0.0378	0.0541	0.0548
e	e	yourself	0.0649	0.0661	0.0588	0.0383	0.4589	0.4774	0.4981	0.5890
ea	eɪ	greatly	0.0078	0.0089	0.0095	0.0089	0.0373	0.0538	0.0714	0.0294
ea	ə	sergeant	0.0078	-	-	-	0.0075	-	-	-
ea	ɪə	theatre	0.0078	0.0089	0.0095	0.0089	0.0896	0.0860	0.0952	0.0882
ea	iː	weakness	0.0078	0.0089	0.0095	0.0089	0.5448	0.5914	0.6190	0.5588
ea	e	widespread	0.0078	0.0089	0.0095	0.0089	0.3060	0.2473	0.1905	0.2941
ea	j	beauty	0.0078	0.0089	0.0095	0.0089	0.0149	0.0215	0.0238	0.0294
ear	ɪə	rear	0.0017	0.0020	0.0039	0.0039	0.5000	0.4762	0.5294	0.4667
ear	eə	wear	0.0017	0.0020	-	0.0039	0.0667	0.0952	-	0.2667
ear	ɑː	heart	0.0017	0.0020	0.0039	-	0.0333	0.0476	0.0588	-

APPENDIX B: (Continued)

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
ear	ɜː	year	0.0017	0.0020	0.0039	0.0039	0.4000	0.3810	0.4118	0.2667
ed	d	transformed	0.0003	0.0002	0.0005	0.0071	0.8000	0.5000	0.5000	0.5185
ed	t	produced	0.0003	0.0002	0.0005	0.0071	0.2000	0.5000	0.5000	0.4815
ee	ɪ	committee	0.0037	0.0042	0.0048	0.0084	0.0156	0.0227	0.0476	0.0313
ee	iː	wheel	0.0037	0.0042	0.0048	0.0084	0.9844	0.9773	0.9524	0.9688
e-e	iː	these	0.0009	0.0012	0.0018	0.0008	0.9375	1.0000	1.0000	1.0000
e-e	e	cigarette	0.0009	-	-	-	0.0625	-	-	-
eer	ɪə	sheer	0.0002	0.0002	-	-	1.0000	1.0000	-	-
ei	aɪ	neither	0.0002	0.0003	0.0005	-	0.5000	0.6667	1.0000	-
ei	iː	receive	0.0002	0.0003	-	-	0.2500	0.3333	-	-
ei	e	leisure	0.0002	-	-	-	0.2500	-	-	-
eig	eɪ	reign	0.0001	-	-	-	0.5000	-	-	-
eig	ə	foreign	0.0001	0.0001	0.0002	-	0.5000	1.0000	1.0000	-
eigh	eɪ	weight	0.0004	0.0004	0.0005	0.0003	0.8571	0.7500	1.0000	1.0000
eigh	aɪ	height	0.0004	0.0004	-	-	0.1429	0.2500	-	-
eir	eə	their	0.0001	0.0001	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
el	l	unfortunately	0.0011	0.0006	0.0005	0.0003	1.0000	1.0000	1.0000	1.0000
en	n	written	0.0011	0.0011	0.0011	0.0008	1.0000	1.0000	1.0000	1.0000
eo	ə	pigeon	-	-	-	0.0005	-	-	-	0.5000
eo	iː	people	0.0001	0.0001	0.0002	0.0005	1.0000	1.0000	1.0000	0.5000
eou	ɪə	simultaneously	0.0001	-	-	-	1.0000	-	-	-
er	ə	younger	0.0163	0.0160	0.0186	0.0147	0.7367	0.7590	0.7683	0.9464
er	ɪə	employer	0.0163	0.0160	-	-	0.0036	0.0060	-	-
er	ɑː	sergeant	0.0163	-	-	-	0.0071	-	-	-
er	ɜː	vertical	0.0163	0.0160	0.0186	0.0147	0.2100	0.1867	0.1707	0.0357
er	ɪ	liberty	0.0163	-	-	-	0.0036	-	-	-
er	r	temperature	0.0163	0.0160	0.0186	0.0147	0.0391	0.0482	0.0610	0.0179
ere	ɪə	sphere	0.0010	0.0013	0.0011	0.0013	0.3333	0.3571	0.2000	0.2000
ere	eə	whereby	0.0010	0.0013	0.0011	0.0013	0.6667	0.6429	0.8000	0.8000
ern	n	pattern	0.0002	0.0002	0.0005	-	1.0000	1.0000	1.0000	-
et	eɪ	ballet	0.0001	-	-	-	1.0000	-	-	-
eu	ɪə	museum	0.0001	0.0001	-	-	1.0000	1.0000	-	-
eur	ə	amateur	0.0001	-	-	-	1.0000	-	-	-
ew	uː	view	0.0012	0.0016	0.0018	0.0037	0.7000	0.7059	0.6250	0.6429
ew	j	newspaper	0.0012	0.0016	0.0018	0.0037	0.3000	0.2941	0.3750	0.3571
ey	eɪ	they	0.0004	0.0006	0.0007	0.0010	0.4286	0.5000	0.6667	0.2500

APPENDIX B: (Continued)

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
ey	iː	valley	0.0004	0.0006	0.0007	0.0010	0.5714	0.5000	0.3333	0.7500
eye	aɪ	eye	0.0001	0.0001	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
f	f	false	0.0176	0.0186	0.0186	0.0215	0.9967	0.9948	0.9878	0.9878
f	v	of	0.0176	0.0186	0.0186	0.0215	0.0033	0.0052	0.0122	0.0122
ff	f	traffic	0.0019	0.0024	0.0025	0.0010	1.0000	1.0000	1.0000	1.0000
ft	f	often	0.0001	0.0001	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
g	dʒ	wage	0.0125	0.0123	0.0109	0.0170	0.2837	0.2656	0.2083	0.1692
g	g	underground	0.0125	0.0123	0.0109	0.0170	0.7163	0.7344	0.7917	0.8308
ge	ʒ	garage	0.0023	-	-	-	0.0256	-	-	-
ge	dʒ	village	0.0023	0.0024	0.0025	0.0016	0.9744	1.0000	1.0000	1.0000
gg	dʒ	suggests	0.0003	0.0006	0.0002	-	0.5000	0.5000	1.0000	-
gg	g	struggle	0.0003	0.0006	-	0.0013	0.5000	0.5000	-	1.0000
gh	f	tough	0.0004	0.0004	0.0002	0.0010	0.8571	1.0000	1.0000	0.7500
gh	g	ghost	0.0004	-	-	0.0010	0.1429	-	-	0.2500
gn	n	sign	0.0002	0.0003	0.0007	0.0005	1.0000	1.0000	1.0000	1.0000
gu	g	guilty	0.0006	0.0006	0.0002	-	1.0000	1.0000	1.0000	-
gue	g	vague	0.0001	0.0001	0.0002	-	1.0000	1.0000	1.0000	-
h	h	unhappy	0.0065	0.0070	0.0104	0.0157	0.9911	1.0000	1.0000	1.0000
h	z	exhaust	0.0065	-	-	-	0.0089	-	-	-
hi	ɪ	vehicle	0.0001	0.0002	-	-	1.0000	1.0000	-	-
ho	ɒ	honour	0.0001	0.0001	-	-	1.0000	1.0000	-	-
hou	aʊ	hours	0.0001	0.0002	0.0005	-	1.0000	1.0000	1.0000	-
i	ə	visible	0.0571	0.0542	0.0482	0.0527	0.0325	0.0408	0.0519	0.0100
i	aɪ	writer	0.0571	0.0542	0.0482	0.0527	0.0832	0.0975	0.0991	0.1095
i	uː	nuisance	0.0571	-	-	-	0.0010	-	-	-
i	ɪ	written	0.0571	0.0542	0.0482	0.0527	0.8772	0.8528	0.8396	0.8756
i	iː	unique	0.0571	0.0542	-	0.0527	0.0030	0.0035	-	0.0050
i	j	view	0.0571	0.0542	0.0482	-	0.0030	0.0053	0.0094	-
ia	ə	parliamentary	0.0010	0.0009	0.0011	-	0.1176	0.2222	0.2000	-
ia	ɪə	variable	0.0010	0.0009	0.0011	0.0003	0.7647	0.6667	0.8000	1.0000
ia	ɪ	marriage	0.0010	0.0009	-	-	0.1176	0.1111	-	-
iar	ɪə	peculiar	0.0001	0.0001	-	-	1.0000	1.0000	-	-
ie	ə	conscience	0.0014	-	-	-	0.0417	-	-	-
ie	ɪə	experience	0.0014	0.0016	0.0016	-	0.1250	0.1176	0.1429	-
ie	aɪ	tie	0.0014	0.0016	-	0.0029	0.1250	0.1176	-	0.1818
ie	iː	yield	0.0014	0.0016	0.0016	0.0029	0.5833	0.6471	0.7143	0.6364

APPENDIX B: (Continued)

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
ie	e	friendship	0.0014	0.0016	0.0016	0.0029	0.1250	0.1176	0.1429	0.1818
i-e	ə	medicine	0.0068	-	-	-	0.0085	-	-	-
i-e	aɪ	write	0.0068	0.0075	0.0075	0.0092	0.8983	0.8846	0.9394	0.9143
i-e	ɪ	imagine	0.0068	0.0075	-	0.0092	0.0593	0.0769	-	0.0286
i-e	iː	routine	0.0068	0.0075	0.0075	0.0092	0.0339	0.0385	0.0606	0.0571
ier	ə	soldier	0.0004	-	-	-	0.1429	-	-	-
ier	ɪə	premier	0.0004	0.0002	0.0002	-	0.8571	1.0000	1.0000	-
igh	aɪ	tonight	0.0014	0.0015	0.0018	0.0021	1.0000	1.0000	1.0000	1.0000
il	l	pupil	0.0004	0.0003	0.0005	0.0005	1.0000	1.0000	1.0000	1.0000
in	n	mountain	0.0003	0.0003	0.0002	0.0005	1.0000	1.0000	1.0000	1.0000
io	ə	transmission	0.0015	0.0018	0.0016	-	0.7308	0.6842	0.5714	-
io	ɪə	union	0.0015	0.0018	0.0016	-	0.2692	0.3158	0.4286	-
ior	ɪə	superior	0.0002	0.0002	0.0002	-	1.0000	1.0000	1.0000	-
iou	ə	unconscious	0.0009	0.0010	-	-	0.3125	0.3000	-	-
iou	ɪə	various	0.0009	0.0010	0.0011	-	0.6875	0.7000	1.0000	-
iour	ɪə	behaviour	0.0001	0.0001	0.0002	-	1.0000	1.0000	1.0000	-
ir	ə	confirmation	0.0012	-	-	-	0.0500	-	-	-
ir	ɜː	virtue	0.0012	0.0011	0.0014	0.0018	0.9500	1.0000	1.0000	1.0000
is	aɪ	isle	0.0001	0.0001	-	-	1.0000	1.0000	-	-
iu	ɪə	medium	0.0002	-	-	-	1.0000	-	-	-
j	dʒ	subject	0.0020	0.0017	0.0014	0.0021	1.0000	1.0000	1.0000	1.0000
k	k	worker	0.0065	0.0065	0.0075	0.0139	1.0000	1.0000	1.0000	1.0000
kn	n	unknown	0.0005	0.0005	0.0009	0.0008	1.0000	1.0000	1.0000	1.0000
l	l	false	0.0431	0.0425	0.0432	0.0433	1.0000	1.0000	1.0000	1.0000
le	l	visible	0.0043	0.0039	0.0036	0.0024	1.0000	1.0000	1.0000	1.0000
ll	l	yellow	0.0057	0.0066	0.0064	0.0094	1.0000	1.0000	1.0000	1.0000
m	m	women	0.0304	0.0309	0.0329	0.0288	1.0000	1.0000	1.0000	1.0000
mb	m	lamb	0.0002	-	-	0.0005	1.0000	-	-	1.0000
mm	m	summit	0.0014	0.0012	0.0014	0.0005	1.0000	1.0000	1.0000	1.0000
mme	m	programme	0.0001	0.0001	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
mn	m	column	0.0001	0.0001	-	-	1.0000	1.0000	-	-
n	ŋ	uncle	0.0769	0.0763	0.0743	0.0558	0.0241	0.0252	0.0245	0.0516
n	n	zone	0.0769	0.0763	0.0743	0.0558	0.9759	0.9748	0.9755	0.9484
nd	n	handsome	0.0001	-	-	-	1.0000	-	-	-
ng	ŋ	younger	0.0030	0.0029	0.0039	0.0136	1.0000	1.0000	1.0000	1.0000
ngue	ŋ	tongue	0.0001	-	-	-	1.0000	-	-	-

APPENDIX B: (Continued)

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
nn	n	winner	0.0008	0.0009	0.0002	0.0013	1.0000	1.0000	1.0000	1.0000
none	w	someone	0.0003	0.0005	0.0011	0.0016	1.0000	1.0000	1.0000	1.0000
o	ə	wisdom	0.0450	0.0430	0.0457	0.0367	0.4736	0.4609	0.3682	0.1714
o	əʊ	zero	0.0450	0.0430	0.0457	0.0367	0.1197	0.1253	0.1542	0.2071
o	ə	woman	0.0450	0.0430	0.0457	-	0.0013	0.0022	0.0050	-
o	ʌ	worry	0.0450	0.0430	0.0457	0.0367	0.0515	0.0559	0.0796	0.1143
o	uː	whose	0.0450	0.0430	0.0457	0.0367	0.0206	0.0291	0.0348	0.0286
o	ɒ	wrong	0.0450	0.0430	0.0457	0.0367	0.3256	0.3199	0.3483	0.4714
o	ɪ	women	0.0450	0.0430	0.0457	-	0.0013	0.0022	0.0050	-
o	ɔː	story	0.0450	0.0430	0.0457	0.0367	0.0064	0.0045	0.0050	0.0071
oa	əʊ	toast	0.0009	0.0011	0.0005	0.0024	0.8000	0.8182	1.0000	1.0000
oa	ɔː	broadcast	0.0009	0.0011	-	-	0.2000	0.1818	-	-
oar	ə	cupboard	0.0001	-	-	-	0.5000	-	-	-
oar	ɔː	board	0.0001	0.0001	0.0002	-	0.5000	1.0000	1.0000	-
oe	əʊ	toe	0.0001	-	-	0.0016	0.5000	-	-	0.5000
oe	ʌ	does	-	-	-	0.0016	-	-	-	0.1667
oe	uː	shoe	0.0001	-	-	0.0016	0.5000	-	-	0.3333
o-e	ə	welcome	0.0043	0.0050	-	0.0087	0.0267	0.0192	-	0.0303
o-e	əʊ	zone	0.0043	0.0050	0.0068	0.0087	0.6133	0.5192	0.4333	0.4242
o-e	ʌ	somewhere	0.0043	0.0050	0.0068	0.0087	0.3333	0.4231	0.5333	0.4242
o-e	ɒ	gone	0.0043	0.0050	0.0068	0.0087	0.0133	0.0192	0.0333	0.1212
o-e	ɔː	moreover	0.0043	0.0050	-	-	0.0133	0.0192	-	-
oi	ə	tortoise	-	-	-	0.0013	-	-	-	0.2000
oi	ɔɪ	voice	0.0008	0.0010	0.0011	0.0013	1.0000	1.0000	1.0000	0.8000
ol	əʊ	folk	0.0001	-	-	-	0.5000	-	-	-
ol	l	symbol	0.0001	-	-	-	0.5000	-	-	-
olo	ɜː	colonel	0.0001	-	-	-	1.0000	-	-	-
on	n	suspicion	0.0008	0.0007	0.0005	0.0003	1.0000	1.0000	1.0000	1.0000
oo	ə	wooden	0.0024	0.0025	0.0027	0.0089	0.4048	0.4615	0.4167	0.3824
oo	ʌ	flood	0.0024	0.0025	0.0027	0.0089	0.0476	0.0385	0.0833	0.0588
oo	uː	tool	0.0024	0.0025	0.0027	0.0089	0.5476	0.5000	0.5000	0.5588
oor	ʊə	poor	0.0002	0.0003	0.0007	0.0010	0.2500	0.3333	0.3333	0.2500
oor	ɔː	floor	0.0002	0.0003	0.0007	0.0010	0.7500	0.6667	0.6667	0.7500
or	ə	visitor	0.0075	0.0075	0.0084	0.0045	0.2403	0.2436	0.1622	0.1176
or	ɜː	worthy	0.0075	0.0075	0.0084	0.0045	0.0853	0.1026	0.1081	0.2353
or	ɔː	worn	0.0075	0.0075	0.0084	0.0045	0.6047	0.5641	0.6757	0.6471

APPENDIX B: (Continued)

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
or	r	theory	0.0075	0.0075	0.0084	-	0.0698	0.0897	0.0541	-
ore	ɔː	wore	0.0008	0.0006	0.0007	0.0005	1.0000	1.0000	1.0000	1.0000
ou	ə	tremendous	0.0051	0.0054	-	0.0079	0.1477	0.1071	-	0.1333
ou	aʊ	without	0.0051	0.0054	0.0064	0.0079	0.5682	0.5893	0.7500	0.7333
ou	ʊə	tourist	0.0051	-	-	-	0.0227	-	-	-
ou	ʌ	younger	0.0051	0.0054	0.0064	0.0079	0.1932	0.2321	0.1786	0.1000
ou	uː	youth	0.0051	0.0054	0.0064	0.0079	0.0682	0.0714	0.0714	0.0333
ough	ə	thoroughly	0.0007	-	-	-	0.1667	-	-	-
ough	əʊ	though	0.0007	0.0010	0.0016	0.0010	0.1667	0.2000	0.2857	0.2500
ough	uː	throughout	0.0007	0.0010	0.0016	0.0010	0.1667	0.2000	0.2857	0.2500
ough	ɔː	thought	0.0007	0.0010	0.0016	0.0010	0.5000	0.6000	0.4286	0.5000
oul	ə	would	0.0002	0.0003	0.0007	0.0008	1.0000	1.0000	1.0000	1.0000
our	ə	neighbourhood	0.0012	0.0012	0.0018	0.0013	0.4286	0.2500	0.2500	0.4000
our	ʊə	tour	0.0012	0.0012	-	-	0.0476	0.0833	-	-
our	ɜː	journey	0.0012	0.0012	-	-	0.1429	0.0833	-	-
our	ɔː	yourself	0.0012	0.0012	0.0018	0.0013	0.3810	0.5833	0.7500	0.6000
ow	aʊ	town	0.0030	0.0041	0.0048	0.0068	0.3462	0.3023	0.3810	0.4615
ow	əʊ	yellow	0.0030	0.0041	0.0048	0.0068	0.6346	0.6744	0.5714	0.5385
ow	ɒ	knowledge	0.0030	0.0041	0.0048	-	0.0192	0.0233	0.0476	-
oy	ɔɪ	unemployment	0.0007	0.0006	0.0007	0.0008	1.0000	1.0000	1.0000	1.0000
p	p	worship	0.0321	0.0323	0.0320	0.0333	1.0000	1.0000	1.0000	1.0000
pb	b	cupboard	0.0001	-	-	-	1.0000	-	-	-
ph	f	triumph	0.0010	0.0009	0.0007	-	1.0000	1.0000	1.0000	-
pp	p	upper	0.0020	0.0019	0.0027	0.0026	1.0000	1.0000	1.0000	1.0000
ps	s	psychology	0.0001	-	-	-	1.0000	-	-	-
q	k	subsequently	0.0020	0.0022	0.0011	0.0018	1.0000	1.0000	1.0000	1.0000
que	k	unique	0.0001	0.0002	-	-	1.0000	1.0000	-	-
r	ə	wire	0.0475	0.0461	0.0386	0.0357	0.0171	0.0230	0.0235	0.0221
r	r	true	0.0475	0.0461	0.0386	0.0357	0.9829	0.9770	0.9765	0.9779
ra	r	extraordinary	0.0001	-	-	-	1.0000	-	-	-
re	ə	figure	0.0001	0.0001	0.0002	-	1.0000	1.0000	1.0000	-
rr	r	worry	0.0019	0.0014	0.0009	0.0021	1.0000	1.0000	1.0000	1.0000
s	ʒ	visual	0.0507	0.0510	0.0513	0.0752	0.0114	0.0113	0.0044	0.0035
s	ʃ	surely	0.0507	0.0510	0.0513	0.0752	0.0057	0.0094	0.0088	0.0070
s	s	yourself	0.0507	0.0510	0.0513	0.0752	0.8331	0.8377	0.8097	0.6028
s	z	wives	0.0507	0.0510	0.0513	0.0752	0.1497	0.1415	0.1770	0.3868

APPENDIX B: (Continued)

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
sc	ʃ	unconscious	0.0005	0.0005	-	-	0.4444	0.2000	-	-
sc	s	scientific	0.0005	0.0005	0.0002	-	0.5556	0.8000	1.0000	-
sch	ʃ	schedule	0.0001	-	-	-	1.0000	-	-	-
se	s	false	0.0028	0.0029	0.0032	0.0029	0.6042	0.5333	0.5000	0.4545
se	z	whose	0.0028	0.0029	0.0032	0.0029	0.3958	0.4667	0.5000	0.5455
sh	ʃ	worship	0.0039	0.0037	0.0030	0.0079	1.0000	1.0000	1.0000	1.0000
si	ʒ	vision	0.0012	0.0012	0.0011	0.0003	0.7000	0.6154	0.8000	1.0000
si	ʃ	version	0.0012	0.0012	0.0011	-	0.3000	0.3846	0.2000	-
ss	ʃ	transmission	0.0059	0.0060	0.0052	-	0.1584	0.1613	0.2174	-
ss	s	witness	0.0059	0.0060	0.0052	0.0029	0.8218	0.8387	0.7826	1.0000
ss	z	possession	0.0059	-	-	-	0.0198	-	-	-
st	s	listen	0.0001	0.0001	-	0.0005	1.0000	1.0000	-	1.0000
t	ʃ	ratio	0.0730	-	-	-	0.0016	-	-	-
t	tʃ	virtue	0.0730	0.0739	0.0734	0.0671	0.0357	0.0338	0.0310	0.0078
t	t	true	0.0730	0.0739	0.0734	0.0671	0.9627	0.9662	0.9690	0.9922
tch	tʃ	watch	0.0005	0.0005	0.0007	0.0013	1.0000	1.0000	1.0000	1.0000
te	t	waste	0.0014	0.0011	0.0007	0.0010	1.0000	1.0000	1.0000	1.0000
th	ð	worthy	0.0077	0.0091	0.0136	0.0123	0.3712	0.3895	0.5333	0.4894
th	θ	youth	0.0077	0.0091	0.0136	0.0123	0.6288	0.6105	0.4667	0.5106
ti	ʒ	equation	0.0094	-	-	-	0.0061	-	-	-
ti	ʃ	variation	0.0094	0.0087	0.0066	0.0005	0.9877	0.9889	0.9655	1.0000
ti	tʃ	question	0.0094	0.0087	0.0066	-	0.0061	0.0111	0.0345	-
tle	l	castle	0.0001	0.0001	-	0.0003	1.0000	1.0000	-	1.0000
tt	t	written	0.0019	0.0020	0.0025	0.0026	1.0000	1.0000	1.0000	1.0000
u	ə	wonderful	0.0270	0.0259	0.0248	0.0202	0.1438	0.1375	0.1468	0.0390
u	ʊə	security	0.0270	0.0259	0.0248	-	0.0150	0.0149	0.0183	-
u	ə	vocabulary	0.0270	0.0259	0.0248	0.0202	0.0880	0.0929	0.0917	0.0779
u	ʌ	utterly	0.0270	0.0259	0.0248	0.0202	0.3369	0.3197	0.3303	0.6494
u	uː	usually	0.0270	0.0259	0.0248	0.0202	0.1395	0.1450	0.1376	0.0649
u	ɪ	minute	0.0270	0.0259	0.0248	0.0202	0.0107	0.0186	0.0275	0.0390
u	j	vocabulary	0.0270	0.0259	0.0248	0.0202	0.1845	0.1822	0.1927	0.0390
u	w	subsequently	0.0270	0.0259	0.0248	0.0202	0.0815	0.0892	0.0550	0.0909
ua	ə	usually	0.0011	0.0013	0.0011	-	0.0526	0.0714	0.2000	-
ua	ʊə	visual	0.0011	0.0013	0.0011	-	0.8947	0.8571	0.6000	-
ua	ə	actually	0.0011	0.0013	0.0011	-	0.0526	0.0714	0.2000	-
ue	ʊə	influence	0.0013	0.0013	0.0018	-	0.0909	0.0714	0.1250	-

APPENDIX B: (Continued)

			Grapheme probability				GPC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Grapheme	Phoneme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
ue	uː	true	0.0013	0.0013	0.0018	0.0008	0.5455	0.5714	0.6250	1.0000
ue	j	value	0.0013	0.0013	0.0018	-	0.3636	0.3571	0.2500	-
u-e	uː	volume	0.0032	0.0027	0.0018	0.0010	0.5636	0.5714	0.6250	0.5000
u-e	j	volume	0.0032	0.0027	0.0018	0.0010	0.4364	0.4286	0.3750	0.5000
ui	uː	suitable	0.0003	0.0003	-	0.0010	0.8000	1.0000	-	0.2500
ui	ɪ	circuit	0.0003	-	-	0.0010	0.2000	-	-	0.7500
ul	l	soul	0.0003	0.0002	-	-	1.0000	1.0000	-	-
ur	ə	survive	0.0017	0.0015	-	-	0.1667	0.1875	-	-
ur	ɜː	urgent	0.0017	0.0015	0.0016	0.0010	0.8333	0.8125	1.0000	1.0000
ure	ə	venture	0.0020	0.0020	0.0018	0.0010	0.8000	0.7619	0.7500	0.7500
ure	ʊə	surely	0.0020	0.0020	0.0018	0.0010	0.2000	0.2381	0.2500	0.2500
uy	aɪ	buy	0.0001	0.0001	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
v	v	wives	0.0117	0.0110	0.0120	0.0073	1.0000	1.0000	1.0000	1.0000
ve	v	twelve	0.0046	0.0053	0.0041	0.0039	1.0000	1.0000	1.0000	1.0000
w	w	wound	0.0070	0.0079	0.0100	0.0152	1.0000	1.0000	1.0000	1.0000
wer	ə	answer	0.0001	0.0001	0.0002	-	1.0000	1.0000	1.0000	-
wh	h	whose	0.0017	0.0021	0.0030	0.0034	0.1724	0.1818	0.3077	0.1538
wh	w	worthwhile	0.0017	0.0021	0.0030	0.0034	0.8276	0.8182	0.6923	0.8462
wo	uː	two	0.0001	0.0001	0.0002	0.0003	1.0000	1.0000	1.0000	1.0000
wr	r	wrote	0.0003	0.0005	0.0009	0.0008	1.0000	1.0000	1.0000	1.0000
x	ʃ	sexual	0.0075	0.0083	-	-	0.0231	0.0233	-	-
x	g	existence	0.0075	0.0083	0.0068	-	0.0692	0.0814	0.0667	-
x	k	text	0.0075	0.0083	0.0068	0.0031	0.4308	0.4186	0.4333	0.5000
x	s	text	0.0075	0.0083	0.0068	0.0031	0.4077	0.3953	0.4333	0.5000
x	z	existence	0.0075	0.0083	0.0068	-	0.0692	0.0814	0.0667	-
y	ə	analysis	0.0250	0.0242	0.0248	-	0.0023	0.0040	0.0092	-
y	aɪ	why	0.0250	0.0242	0.0248	0.0191	0.0648	0.0635	0.0550	0.1507
y	ɪ	typical	0.0250	0.0242	0.0248	0.0191	0.0278	0.0119	0.0183	0.1096
y	iː	worthy	0.0250	0.0242	0.0248	0.0191	0.8681	0.8690	0.8349	0.6301
y	j	youth	0.0250	0.0242	0.0248	0.0191	0.0370	0.0516	0.0826	0.1096
y-e	aɪ	type	0.0001	0.0002	0.0005	-	1.0000	1.0000	1.0000	-
z	s	pizza	-	-	-	0.0016	-	-	-	0.1667
z	t	pizza	-	-	-	0.0016	-	-	-	0.1667
z	z	zone	0.0008	0.0004	0.0002	0.0016	1.0000	1.0000	1.0000	0.6667
ze	z	bronze	0.0001	-	-	-	1.0000	-	-	-

APPENDIX C: Sonograph and PGC Probability

Sonograph probability = Sonograph probability of occurrence. PGC probability = Phoneme-grapheme correspondence (spelling) probability.

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
æ	a	as	0.0188	0.0175	0.0164	0.0218	1.0000	1.0000	1.0000	1.0000
ɑː	a	vast	0.0032	0.0037	0.0039	0.0042	0.4231	0.4368	0.4146	0.4211
ɑː	ar	yard	0.0038	0.0042	0.0048	0.0045	0.5000	0.5057	0.5122	0.4474
ɑː	are	are	0.0001	0.0001	0.0002	0.0003	0.0077	0.0115	0.0244	0.0263
ɑː	au	laughter	0.0003	0.0003	0.0002	0.0010	0.0462	0.0345	0.0244	0.1053
ɑː	ear	heart	0.0001	0.0001	0.0002	-	0.0077	0.0115	0.0244	-
ɑː	er	sergeant	0.0001	-	-	-	0.0154	-	-	-
b	b	whereby	0.0177	0.0173	0.0179	0.0270	0.9903	1.0000	1.0000	0.9626
b	bb	rubber	0.0001	-	-	0.0010	0.0065	-	-	0.0374
b	pb	cupboard	0.0001	-	-	-	0.0032	-	-	-
k	c	welcome	0.0315	0.0309	0.0270	0.0220	0.6618	0.6538	0.6611	0.4719
k	cc	occurrence	0.0009	0.0005	0.0002	0.0003	0.0182	0.0102	0.0056	0.0056
k	ch	technology	0.0011	0.0009	0.0009	0.0010	0.0231	0.0183	0.0222	0.0225
k	ck	wicked	0.0022	0.0026	0.0011	0.0060	0.0462	0.0550	0.0278	0.1292
k	cq	acquisition	0.0001	-	-	-	0.0024	-	-	-
k	k	worker	0.0065	0.0065	0.0075	0.0139	0.1363	0.1385	0.1833	0.2978
k	q	subsequently	0.0020	0.0022	0.0011	0.0018	0.0414	0.0468	0.0278	0.0393
k	que	unique	0.0001	0.0002	-	-	0.0024	0.0041	-	-
k	x	text	0.0032	0.0035	0.0030	0.0016	0.0681	0.0733	0.0722	0.0337
tʃ	ch	which	0.0035	0.0043	0.0050	0.0055	0.5217	0.5844	0.6111	0.7500
tʃ	t	virtue	0.0026	0.0025	0.0023	0.0005	0.3913	0.3377	0.2778	0.0714
tʃ	tch	watch	0.0005	0.0005	0.0007	0.0013	0.0783	0.0649	0.0833	0.1786
tʃ	ti	question	0.0001	0.0001	0.0002	-	0.0087	0.0130	0.0278	-
d	d	yield	0.0354	0.0357	0.0375	0.0414	0.9745	0.9712	0.9763	0.9080
d	dd	wedding	0.0007	0.0010	0.0007	0.0005	0.0191	0.0262	0.0178	0.0115
d	ed	transformed	0.0002	0.0001	0.0002	0.0037	0.0064	0.0026	0.0059	0.0805
iː	ae	aesthetic	0.0001	-	-	-	0.0017	-	-	-
iː	e	we	0.0024	0.0025	0.0032	0.0021	0.0705	0.0690	0.0833	0.0661
iː	ea	weakness	0.0042	0.0053	0.0059	0.0050	0.1225	0.1459	0.1548	0.1570
iː	ee	wheel	0.0037	0.0041	0.0045	0.0081	0.1057	0.1141	0.1190	0.2562
iː	e-e	these	0.0009	0.0012	0.0018	0.0008	0.0252	0.0345	0.0476	0.0248
iː	ei	receive	0.0001	0.0001	-	-	0.0017	0.0027	-	-
iː	eo	people	0.0001	0.0001	0.0002	0.0003	0.0017	0.0027	0.0060	0.0083
iː	ey	valley	0.0002	0.0003	0.0002	0.0008	0.0067	0.0080	0.0060	0.0248
iː	i	unique	0.0002	0.0002	-	0.0003	0.0050	0.0053	-	0.0083
iː	ie	yield	0.0008	0.0011	0.0011	0.0018	0.0235	0.0292	0.0298	0.0579

APPENDIX C: (Continued)

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
iː	i-e	routine	0.0002	0.0003	0.0005	0.0005	0.0067	0.0080	0.0119	0.0165
iː	y	worthy	0.0217	0.0211	0.0207	0.0121	0.6292	0.5809	0.5417	0.3802
ə	a	woman	0.0224	0.0197	0.0173	0.0081	0.2622	0.2452	0.2331	0.2053
ə	ai	uncertainty	0.0003	0.0002	0.0005	0.0003	0.0034	0.0024	0.0061	0.0066
ə	ar	upwards	0.0014	0.0014	0.0020	0.0005	0.0169	0.0179	0.0276	0.0132
ə	au	restaurant	0.0001	0.0001	-	-	0.0007	0.0012	-	-
ə	e	witness	0.0140	0.0134	0.0107	0.0047	0.1640	0.1663	0.1442	0.1192
ə	ea	sergeant	0.0001	-	-	-	0.0007	-	-	-
ə	eig	foreign	0.0001	0.0001	0.0002	-	0.0007	0.0012	0.0031	-
ə	eo	pigeon	-	-	-	0.0003	-	-	-	0.0066
ə	er	younger	0.0120	0.0121	0.0143	0.0139	0.1402	0.1507	0.1933	0.3510
ə	eur	amateur	0.0001	-	-	-	0.0007	-	-	-
ə	i	visible	0.0019	0.0022	0.0025	0.0005	0.0217	0.0275	0.0337	0.0132
ə	ia	parliamentary	0.0001	0.0002	0.0002	-	0.0014	0.0024	0.0031	-
ə	ie	conscience	0.0001	-	-	-	0.0007	-	-	-
ə	i-e	medicine	0.0001	-	-	-	0.0007	-	-	-
ə	ier	soldier	0.0001	-	-	-	0.0007	-	-	-
ə	io	transmission	0.0011	0.0012	0.0009	-	0.0129	0.0156	0.0123	-
ə	iou	unconscious	0.0003	0.0003	-	-	0.0034	0.0036	-	-
ə	ir	confirmation	0.0001	-	-	-	0.0007	-	-	-
ə	o	wisdom	0.0213	0.0198	0.0168	0.0063	0.2493	0.2464	0.2270	0.1589
ə	oar	cupboard	0.0001	-	-	-	0.0007	-	-	-
ə	o-e	welcome	0.0001	0.0001	-	0.0003	0.0014	0.0012	-	0.0066
ə	oi	tortoise	-	-	-	0.0003	-	-	-	0.0066
ə	or	visitor	0.0018	0.0018	0.0014	0.0005	0.0210	0.0227	0.0184	0.0132
ə	ou	tremendous	0.0008	0.0006	-	0.0010	0.0088	0.0072	-	0.0265
ə	ough	thoroughly	0.0001	-	-	-	0.0014	-	-	-
ə	our	neighbourhood	0.0005	0.0003	0.0005	0.0005	0.0061	0.0036	0.0061	0.0132
ə	r	wire	0.0008	0.0011	0.0009	0.0008	0.0095	0.0132	0.0123	0.0199
ə	re	figure	0.0001	0.0001	0.0002	-	0.0007	0.0012	0.0031	-
ə	u	wonderful	0.0039	0.0036	0.0036	0.0008	0.0454	0.0443	0.0491	0.0199
ə	ua	usually	0.0001	0.0001	0.0002	-	0.0007	0.0012	0.0031	-
ə	ur	survive	0.0003	0.0003	-	-	0.0034	0.0036	-	-
ə	ure	venture	0.0016	0.0015	0.0014	0.0008	0.0190	0.0191	0.0184	0.0199
ə	wer	answer	0.0001	0.0001	0.0002	-	0.0007	0.0012	0.0031	-
ə	y	analysis	0.0001	0.0001	0.0002	-	0.0007	0.0012	0.0031	-

APPENDIX C: (Continued)

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
e	a	necessarily	0.0005	0.0008	0.0011	0.0010	0.0140	0.0220	0.0342	0.0385
e	ai	said	0.0002	0.0003	0.0007	0.0005	0.0053	0.0082	0.0205	0.0192
e	e	yourself	0.0298	0.0315	0.0293	0.0225	0.9002	0.9011	0.8836	0.8269
e	ea	widespread	0.0024	0.0022	0.0018	0.0026	0.0718	0.0632	0.0548	0.0962
e	e-e	cigarette	0.0001	-	-	-	0.0018	-	-	-
e	ei	leisure	0.0001	-	-	-	0.0018	-	-	-
e	ie	friendship	0.0002	0.0002	0.0002	0.0005	0.0053	0.0055	0.0068	0.0192
ɜː	ear	year	0.0007	0.0008	0.0016	0.0010	0.0923	0.1111	0.1842	0.1905
ɜː	er	vertical	0.0034	0.0030	0.0032	0.0005	0.4538	0.4306	0.3684	0.0952
ɜː	ir	virtue	0.0011	0.0011	0.0014	0.0018	0.1462	0.1528	0.1579	0.3333
ɜː	olo	colonel	0.0001	-	-	-	0.0077	-	-	-
ɜː	or	worthy	0.0006	0.0008	0.0009	0.0010	0.0846	0.1111	0.1053	0.1905
ɜː	our	journey	0.0002	0.0001	-	-	0.0231	0.0139	-	-
ɜː	ur	urgent	0.0014	0.0012	0.0016	0.0010	0.1923	0.1806	0.1842	0.1905
ɪə	e	zero	0.0009	0.0009	0.0014	0.0003	0.1495	0.1429	0.1667	0.0769
ɪə	ea	theatre	0.0007	0.0008	0.0009	0.0008	0.1121	0.1270	0.1111	0.2308
ɪə	ear	rear	0.0009	0.0010	0.0020	0.0018	0.1402	0.1587	0.2500	0.5385
ɪə	eer	sheer	0.0002	0.0002	-	-	0.0374	0.0317	-	-
ɪə	eou	simultaneously	0.0001	-	-	-	0.0093	-	-	-
ɪə	er	employer	0.0001	0.0001	-	-	0.0093	0.0159	-	-
ɪə	ere	sphere	0.0003	0.0005	0.0002	0.0003	0.0561	0.0794	0.0278	0.0769
ɪə	eu	museum	0.0001	0.0001	-	-	0.0093	0.0159	-	-
ɪə	ia	variable	0.0008	0.0006	0.0009	0.0003	0.1215	0.0952	0.1111	0.0769
ɪə	iar	peculiar	0.0001	0.0001	-	-	0.0187	0.0159	-	-
ɪə	ie	experience	0.0002	0.0002	0.0002	-	0.0280	0.0317	0.0278	-
ɪə	ier	premier	0.0003	0.0002	0.0002	-	0.0561	0.0317	0.0278	-
ɪə	io	union	0.0004	0.0006	0.0007	-	0.0654	0.0952	0.0833	-
ɪə	ior	superior	0.0002	0.0002	0.0002	-	0.0374	0.0317	0.0278	-
ɪə	iou	various	0.0006	0.0007	0.0011	-	0.1028	0.1111	0.1389	-
ɪə	iour	behaviour	0.0001	0.0001	0.0002	-	0.0093	0.0159	0.0278	-
ɪə	iu	medium	0.0002	-	-	-	0.0374	-	-	-
eə	a	vary	0.0004	0.0004	0.0005	-	0.1129	0.1053	0.1333	-
eə	ai	dairy	0.0001	-	-	0.0003	0.0161	-	-	0.0526
eə	air	upstairs	0.0009	0.0009	0.0009	0.0013	0.2419	0.2368	0.2667	0.2632
eə	ar	scarcely	0.0001	-	-	-	0.0161	-	-	-
eə	are	welfare	0.0012	0.0012	0.0009	0.0010	0.3226	0.3421	0.2667	0.2105

APPENDIX C: (Continued)

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
eə	ayer	prayer	0.0001	-	-	-	0.0161	-	-	-
eə	ayor	mayor	0.0001	-	-	-	0.0161	-	-	-
eə	e	wherever	0.0001	-	-	-	0.0161	-	-	-
eə	ear	wear	0.0001	0.0002	-	0.0010	0.0323	0.0526	-	0.2105
eə	eir	their	0.0001	0.0001	0.0002	0.0003	0.0161	0.0263	0.0667	0.0526
eə	ere	whereby	0.0007	0.0009	0.0009	0.0010	0.1935	0.2368	0.2667	0.2105
eɪ	a	waste	0.0079	0.0072	0.0068	0.0050	0.4114	0.3731	0.3333	0.2436
eɪ	a-e	wave	0.0056	0.0058	0.0066	0.0073	0.2913	0.2985	0.3222	0.3590
eɪ	ai	wait	0.0030	0.0033	0.0020	0.0024	0.1562	0.1692	0.1000	0.1154
eɪ	aigh	straight	0.0001	0.0001	0.0002	0.0003	0.0030	0.0050	0.0111	0.0128
eɪ	ay	yesterday	0.0017	0.0019	0.0032	0.0047	0.0901	0.0995	0.1556	0.2308
eɪ	ea	greatly	0.0003	0.0005	0.0007	0.0003	0.0150	0.0249	0.0333	0.0128
eɪ	eig	reign	0.0001	-	-	-	0.0030	-	-	-
eɪ	eigh	weight	0.0003	0.0003	0.0005	0.0003	0.0180	0.0149	0.0222	0.0128
eɪ	et	ballet	0.0001	-	-	-	0.0030	-	-	-
eɪ	ey	they	0.0002	0.0003	0.0005	0.0003	0.0090	0.0149	0.0222	0.0128
f	f	false	0.0176	0.0185	0.0184	0.0212	0.8440	0.8312	0.8351	0.9101
f	ff	traffic	0.0019	0.0024	0.0025	0.0010	0.0891	0.1082	0.1134	0.0449
f	ft	often	0.0001	0.0001	0.0002	0.0003	0.0028	0.0043	0.0103	0.0112
f	gh	tough	0.0003	0.0004	0.0002	0.0008	0.0167	0.0173	0.0103	0.0337
f	ph	triumph	0.0010	0.0009	0.0007	-	0.0474	0.0390	0.0309	-
g	g	underground	0.0089	0.0090	0.0086	0.0142	0.8603	0.8468	0.9048	0.9000
g	gg	struggle	0.0002	0.0003	-	0.0013	0.0168	0.0270	-	0.0833
g	gh	ghost	0.0001	-	-	0.0003	0.0056	-	-	0.0167
g	gu	guilty	0.0006	0.0006	0.0002	-	0.0559	0.0541	0.0238	-
g	gue	vague	0.0001	0.0001	0.0002	-	0.0112	0.0090	0.0238	-
g	x	existence	0.0005	0.0007	0.0005	-	0.0503	0.0631	0.0476	-
h	h	unhappy	0.0064	0.0070	0.0104	0.0157	0.9569	0.9481	0.9200	0.9677
h	wh	whose	0.0003	0.0004	0.0009	0.0005	0.0431	0.0519	0.0800	0.0323
ɪ	a	village	0.0014	0.0019	0.0018	0.0010	0.0204	0.0284	0.0313	0.0175
ɪ	e	women	0.0177	0.0178	0.0143	0.0087	0.2488	0.2624	0.2461	0.1441
ɪ	ee	committee	0.0001	0.0001	0.0002	0.0003	0.0008	0.0014	0.0039	0.0044
ɪ	er	liberty	0.0001	-	-	-	0.0008	-	-	-
ɪ	hi	vehicle	0.0001	0.0002	-	-	0.0016	0.0028	-	-
ɪ	i	written	0.0501	0.0462	0.0404	0.0461	0.7047	0.6823	0.6953	0.7686
ɪ	ia	marriage	0.0001	0.0001	-	-	0.0016	0.0014	-	-

APPENDIX C: (Continued)

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
ɪ	i-e	imagine	0.0004	0.0006	-	0.0003	0.0057	0.0085	-	0.0044
ɪ	o	women	0.0001	0.0001	0.0002	-	0.0008	0.0014	0.0039	-
ɪ	u	minute	0.0003	0.0005	0.0007	0.0008	0.0041	0.0071	0.0117	0.0131
ɪ	ui	circuit	0.0001	-	-	0.0008	0.0008	-	-	0.0131
ɪ	y	typical	0.0007	0.0003	0.0005	0.0021	0.0098	0.0043	0.0078	0.0349
aɪ	ei	neither	0.0001	0.0002	0.0005	-	0.0079	0.0120	0.0278	-
aɪ	eigh	height	0.0001	0.0001	-	-	0.0040	0.0060	-	-
aɪ	eye	eye	0.0001	0.0001	0.0002	0.0003	0.0040	0.0060	0.0139	0.0130
aɪ	i	writer	0.0048	0.0053	0.0048	0.0058	0.3254	0.3313	0.2917	0.2857
aɪ	ie	tie	0.0002	0.0002	-	0.0005	0.0119	0.0120	-	0.0260
aɪ	i-e	write	0.0061	0.0066	0.0070	0.0084	0.4206	0.4157	0.4306	0.4156
aɪ	igh	tonight	0.0014	0.0015	0.0018	0.0021	0.0952	0.0964	0.1111	0.1039
aɪ	is	isle	0.0001	0.0001	-	-	0.0079	0.0060	-	-
aɪ	uy	buy	0.0001	0.0001	0.0002	0.0003	0.0040	0.0060	0.0139	0.0130
aɪ	y	why	0.0016	0.0015	0.0014	0.0029	0.1111	0.0964	0.0833	0.1429
aɪ	y-e	type	0.0001	0.0002	0.0005	-	0.0079	0.0120	0.0278	-
dʒ	d	soldier	0.0003	0.0005	0.0005	-	0.0390	0.0549	0.0625	-
dʒ	dg	judgment	0.0001	0.0002	0.0002	0.0003	0.0130	0.0220	0.0313	0.0345
dʒ	dge	ridge	0.0005	0.0004	0.0002	0.0008	0.0519	0.0440	0.0313	0.1034
dʒ	dj	adjustment	0.0001	-	-	-	0.0065	-	-	-
dʒ	g	wage	0.0035	0.0033	0.0023	0.0029	0.3961	0.3736	0.3125	0.3793
dʒ	ge	village	0.0022	0.0024	0.0025	0.0016	0.2468	0.2747	0.3438	0.2069
dʒ	gg	suggests	0.0002	0.0003	0.0002	-	0.0195	0.0330	0.0313	-
dʒ	j	subject	0.0020	0.0017	0.0014	0.0021	0.2273	0.1978	0.1875	0.2759
l	al	vital	0.0046	0.0049	0.0043	0.0008	0.0777	0.0829	0.0739	0.0138
l	el	unfortunately	0.0011	0.0006	0.0005	0.0003	0.0185	0.0098	0.0078	0.0046
l	il	pupil	0.0004	0.0003	0.0005	0.0005	0.0068	0.0049	0.0078	0.0092
l	l	false	0.0431	0.0425	0.0432	0.0433	0.7221	0.7187	0.7393	0.7604
l	le	visible	0.0043	0.0039	0.0036	0.0024	0.0719	0.0667	0.0623	0.0415
l	ll	yellow	0.0057	0.0066	0.0064	0.0094	0.0962	0.1122	0.1089	0.1659
l	ol	symbol	0.0001	-	-	-	0.0010	-	-	-
l	tle	castle	0.0001	0.0001	-	0.0003	0.0010	0.0016	-	0.0046
l	ul	soul	0.0003	0.0002	-	-	0.0049	0.0033	-	-
m	m	women	0.0304	0.0309	0.0329	0.0288	0.9459	0.9554	0.9539	0.9565
m	mb	lamb	0.0002	-	-	0.0005	0.0054	-	-	0.0174
m	mm	summit	0.0014	0.0012	0.0014	0.0005	0.0432	0.0387	0.0395	0.0174

APPENDIX C: (Continued)

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
m	mme	programme	0.0001	0.0001	0.0002	0.0003	0.0018	0.0030	0.0066	0.0087
m	mn	column	0.0001	0.0001	-	-	0.0036	0.0030	-	-
ŋ	n	uncle	0.0019	0.0019	0.0018	0.0029	0.3765	0.4000	0.3200	0.1746
ŋ	ng	younger	0.0030	0.0029	0.0039	0.0136	0.6118	0.6000	0.6800	0.8254
ŋ	ngue	tongue	0.0001	-	-	-	0.0118	-	-	-
n	en	written	0.0011	0.0011	0.0011	0.0008	0.0139	0.0135	0.0148	0.0138
n	ern	pattern	0.0002	0.0002	0.0005	-	0.0029	0.0025	0.0059	-
n	gn	sign	0.0002	0.0003	0.0007	0.0005	0.0029	0.0037	0.0089	0.0092
n	in	mountain	0.0003	0.0003	0.0002	0.0005	0.0037	0.0037	0.0030	0.0092
n	kn	unknown	0.0005	0.0005	0.0009	0.0008	0.0059	0.0061	0.0119	0.0138
n	n	zone	0.0751	0.0744	0.0725	0.0530	0.9494	0.9509	0.9466	0.9266
n	nd	handsome	0.0001	-	-	-	0.0015	-	-	-
n	nn	winner	0.0008	0.0009	0.0002	0.0013	0.0095	0.0111	0.0030	0.0229
n	on	suspicion	0.0008	0.0007	0.0005	0.0003	0.0103	0.0086	0.0059	0.0046
ɒ	a	what	0.0008	0.0009	0.0016	0.0031	0.0478	0.0577	0.0875	0.1446
ɒ	ach	yacht	0.0001	-	-	-	0.0037	-	-	-
ɒ	au	because	0.0001	0.0001	0.0002	0.0003	0.0037	0.0064	0.0125	0.0120
ɒ	ho	honour	0.0001	0.0001	-	-	0.0074	0.0064	-	-
ɒ	o	wrong	0.0147	0.0137	0.0159	0.0173	0.9301	0.9167	0.8750	0.7952
ɒ	o-e	gone	0.0001	0.0001	0.0002	0.0010	0.0037	0.0064	0.0125	0.0482
ɒ	ow	knowledge	0.0001	0.0001	0.0002	-	0.0037	0.0064	0.0125	-
əʊ	o	zero	0.0054	0.0054	0.0070	0.0076	0.4947	0.4553	0.5167	0.4143
əʊ	oa	toast	0.0007	0.0009	0.0005	0.0024	0.0638	0.0732	0.0333	0.1286
əʊ	oe	toe	0.0001	-	-	0.0008	0.0053	-	-	0.0429
əʊ	o-e	zone	0.0027	0.0026	0.0030	0.0037	0.2447	0.2195	0.2167	0.2000
əʊ	ol	folk	0.0001	-	-	-	0.0053	-	-	-
əʊ	ough	though	0.0001	0.0002	0.0005	0.0003	0.0106	0.0163	0.0333	0.0143
əʊ	ow	yellow	0.0019	0.0028	0.0027	0.0037	0.1755	0.2358	0.2000	0.2000
uː	ew	view	0.0008	0.0012	0.0011	0.0024	0.0795	0.1081	0.1042	0.1875
uː	i	nuisance	0.0001	-	-	-	0.0057	-	-	-
uː	o	whose	0.0009	0.0012	0.0016	0.0010	0.0909	0.1171	0.1458	0.0833
uː	oe	shoe	0.0001	-	-	0.0005	0.0057	-	-	0.0417
uː	oo	tool	0.0013	0.0012	0.0014	0.0050	0.1307	0.1171	0.1250	0.3958
uː	ou	youth	0.0003	0.0004	0.0005	0.0003	0.0341	0.0360	0.0417	0.0208
uː	ough	throughout	0.0001	0.0002	0.0005	0.0003	0.0114	0.0180	0.0417	0.0208
uː	u	usually	0.0038	0.0037	0.0034	0.0013	0.3693	0.3514	0.3125	0.1042

APPENDIX C: (Continued)

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
uː	ue	true	0.0007	0.0008	0.0011	0.0008	0.0682	0.0721	0.1042	0.0625
uː	u-e	volume	0.0018	0.0015	0.0011	0.0005	0.1761	0.1441	0.1042	0.0417
uː	ui	suitable	0.0002	0.0003	-	0.0003	0.0227	0.0270	-	0.0208
uː	wo	two	0.0001	0.0001	0.0002	0.0003	0.0057	0.0090	0.0208	0.0208
ɔɪ	oi	voice	0.0008	0.0010	0.0011	0.0010	0.5200	0.6250	0.6250	0.5714
ɔɪ	oy	unemployment	0.0007	0.0006	0.0007	0.0008	0.4800	0.3750	0.3750	0.4286
ʊə	oor	poor	0.0001	0.0001	0.0002	0.0003	0.0270	0.0417	0.1111	0.5000
ʊə	ou	tourist	0.0001	-	-	-	0.0541	-	-	-
ʊə	our	tour	0.0001	0.0001	-	-	0.0270	0.0417	-	-
ʊə	u	security	0.0004	0.0004	0.0005	-	0.1892	0.1667	0.2222	-
ʊə	ua	visual	0.0010	0.0012	0.0007	-	0.4595	0.5000	0.3333	-
ʊə	ue	influence	0.0001	0.0001	0.0002	-	0.0541	0.0417	0.1111	-
ʊə	ure	surely	0.0004	0.0005	0.0005	0.0003	0.1892	0.2083	0.2222	0.5000
ɔː	a	false	0.0013	0.0020	0.0030	0.0039	0.1292	0.1826	0.2031	0.3061
ɔː	al	walk	0.0001	0.0002	0.0005	0.0008	0.0112	0.0174	0.0313	0.0612
ɔː	ar	warmth	0.0004	0.0005	0.0005	0.0005	0.0393	0.0435	0.0313	0.0408
ɔː	au	pause	0.0007	0.0008	0.0005	0.0003	0.0674	0.0696	0.0313	0.0204
ɔː	augh	taught	0.0002	0.0003	0.0005	0.0005	0.0169	0.0261	0.0313	0.0408
ɔː	aur	dinosaur	-	-	-	0.0003	-	-	-	0.0204
ɔː	aw	withdrawn	0.0007	0.0005	0.0005	0.0008	0.0674	0.0435	0.0313	0.0612
ɔː	o	story	0.0003	0.0002	0.0002	0.0003	0.0281	0.0174	0.0156	0.0204
ɔː	oa	broadcast	0.0002	0.0002	-	-	0.0169	0.0174	-	-
ɔː	oar	board	0.0001	0.0001	0.0002	-	0.0056	0.0087	0.0156	-
ɔː	o-e	moreover	0.0001	0.0001	-	-	0.0056	0.0087	-	-
ɔː	oor	floor	0.0002	0.0002	0.0005	0.0008	0.0169	0.0174	0.0313	0.0612
ɔː	or	worn	0.0045	0.0042	0.0057	0.0029	0.4382	0.3826	0.3906	0.2245
ɔː	ore	wore	0.0008	0.0006	0.0007	0.0005	0.0787	0.0522	0.0469	0.0408
ɔː	ough	thought	0.0003	0.0006	0.0007	0.0005	0.0337	0.0522	0.0469	0.0408
ɔː	our	yourself	0.0005	0.0007	0.0014	0.0008	0.0449	0.0609	0.0938	0.0612
aʊ	hou	hours	0.0001	0.0002	0.0005	-	0.0286	0.0417	0.0645	-
aʊ	ou	without	0.0029	0.0032	0.0048	0.0058	0.7143	0.6875	0.6774	0.6471
aʊ	ow	town	0.0010	0.0012	0.0018	0.0031	0.2571	0.2708	0.2581	0.3529
p	p	worship	0.0321	0.0323	0.0320	0.0333	0.9421	0.9438	0.9216	0.9270
p	pp	upper	0.0020	0.0019	0.0027	0.0026	0.0579	0.0562	0.0784	0.0730
r	er	temperature	0.0006	0.0008	0.0011	0.0003	0.0127	0.0159	0.0276	0.0069
r	or	theory	0.0005	0.0007	0.0005	-	0.0104	0.0139	0.0110	-

APPENDIX C: (Continued)

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
r	r	true	0.0467	0.0450	0.0377	0.0349	0.9307	0.9304	0.9171	0.9172
r	ra	extraordinary	0.0001	-	-	-	0.0012	-	-	-
r	rr	worry	0.0019	0.0014	0.0009	0.0021	0.0381	0.0298	0.0221	0.0552
r	wr	wrote	0.0003	0.0005	0.0009	0.0008	0.0069	0.0099	0.0221	0.0207
ʒ	ge	garage	0.0001	-	-	-	0.0385	-	-	-
ʒ	s	visual	0.0006	0.0006	0.0002	0.0003	0.3846	0.4286	0.2000	0.5000
ʒ	si	vision	0.0008	0.0008	0.0009	0.0003	0.5385	0.5714	0.8000	0.5000
ʒ	ti	equation	0.0001	-	-	-	0.0385	-	-	-
s	c	velocity	0.0071	0.0067	0.0080	0.0050	0.1106	0.1036	0.1241	0.0856
s	ce	voice	0.0050	0.0053	0.0057	0.0013	0.0782	0.0814	0.0887	0.0225
s	ps	psychology	0.0001	-	-	-	0.0018	-	-	-
s	s	yourself	0.0423	0.0427	0.0416	0.0454	0.6556	0.6568	0.6489	0.7793
s	sc	scientific	0.0003	0.0004	0.0002	-	0.0045	0.0059	0.0035	-
s	se	false	0.0017	0.0015	0.0016	0.0013	0.0261	0.0237	0.0248	0.0225
s	ss	witness	0.0048	0.0050	0.0041	0.0029	0.0746	0.0769	0.0638	0.0495
s	st	listen	0.0001	0.0001	-	0.0005	0.0009	0.0015	-	0.0090
s	x	text	0.0031	0.0033	0.0030	0.0016	0.0477	0.0503	0.0461	0.0270
s	z	pizza	-	-	-	0.0003	-	-	-	0.0045
z	h	exhaust	0.0001	-	-	-	0.0056	-	-	-
z	s	wives	0.0076	0.0072	0.0091	0.0291	0.7360	0.7500	0.8000	0.9174
z	se	whose	0.0011	0.0013	0.0016	0.0016	0.1067	0.1400	0.1400	0.0496
z	ss	possession	0.0001	-	-	-	0.0112	-	-	-
z	x	existence	0.0005	0.0007	0.0005	-	0.0506	0.0700	0.0400	-
z	z	zone	0.0008	0.0004	0.0002	0.0010	0.0787	0.0400	0.0200	0.0331
z	ze	bronze	0.0001	-	-	-	0.0112	-	-	-
ʃ	c	appreciation	0.0001	-	-	-	0.0069	-	-	-
ʃ	ch	machinery	0.0001	0.0001	0.0002	0.0003	0.0069	0.0061	0.0182	0.0278
ʃ	ci	suspicion	0.0012	0.0012	0.0011	0.0003	0.0724	0.0736	0.0909	0.0278
ʃ	s	surely	0.0003	0.0005	0.0005	0.0005	0.0172	0.0307	0.0364	0.0556
ʃ	sc	unconscious	0.0002	0.0001	-	-	0.0138	0.0061	-	-
ʃ	sch	schedule	0.0001	-	-	-	0.0034	-	-	-
ʃ	sh	worship	0.0039	0.0037	0.0030	0.0079	0.2310	0.2331	0.2364	0.8333
ʃ	si	version	0.0003	0.0005	0.0002	-	0.0207	0.0307	0.0182	-
ʃ	ss	transmission	0.0009	0.0010	0.0011	-	0.0552	0.0613	0.0909	-
ʃ	t	ratio	0.0001	-	-	-	0.0069	-	-	-
ʃ	ti	variation	0.0093	0.0086	0.0064	0.0005	0.5552	0.5460	0.5091	0.0556

APPENDIX C: (Continued)

			Sonograph probability				PGC probability			
			Corpus A			Corpus C	Corpus A			Corpus C
Phoneme	Grapheme	Example Word	3K	2K	1K	1K	3K	2K	1K	1K
ʃ	x	sexual	0.0002	0.0002	-	-	0.0103	0.0123	-	-
t	bt	undoubtedly	0.0003	0.0002	0.0002	-	0.0039	0.0026	0.0030	-
t	d	chased	-	-	-	0.0005	-	-	-	0.0070
t	ed	produced	0.0001	0.0001	0.0002	0.0034	0.0008	0.0013	0.0030	0.0458
t	t	true	0.0703	0.0714	0.0711	0.0666	0.9506	0.9550	0.9514	0.8944
t	te	waste	0.0014	0.0011	0.0007	0.0010	0.0196	0.0141	0.0091	0.0141
t	tt	written	0.0019	0.0020	0.0025	0.0026	0.0251	0.0270	0.0334	0.0352
t	z	pizza	-	-	-	0.0003	-	-	-	0.0035
ð	th	worthy	0.0028	0.0036	0.0073	0.0060	1.0000	1.0000	1.0000	1.0000
θ	th	youth	0.0048	0.0056	0.0064	0.0063	1.0000	1.0000	1.0000	1.0000
ə	o	woman	0.0001	0.0001	0.0002	-	0.0159	0.0238	0.0500	-
ə	oo	wooden	0.0010	0.0012	0.0011	0.0034	0.2698	0.2857	0.2500	0.5909
ə	oul	would	0.0002	0.0003	0.0007	0.0008	0.0476	0.0714	0.1500	0.1364
ə	u	vocabulary	0.0024	0.0024	0.0023	0.0016	0.6508	0.5952	0.5000	0.2727
ə	ua	actually	0.0001	0.0001	0.0002	-	0.0159	0.0238	0.0500	-
ʌ	o	worry	0.0023	0.0024	0.0036	0.0042	0.1660	0.1701	0.2162	0.1860
ʌ	oe	does	-	-	-	0.0003	-	-	-	0.0116
ʌ	o-e	somewhere	0.0014	0.0021	0.0036	0.0037	0.1037	0.1497	0.2162	0.1628
ʌ	oo	flood	0.0001	0.0001	0.0002	0.0005	0.0083	0.0068	0.0135	0.0233
ʌ	ou	younger	0.0010	0.0012	0.0011	0.0008	0.0705	0.0884	0.0676	0.0349
ʌ	u	utterly	0.0091	0.0083	0.0082	0.0131	0.6515	0.5850	0.4865	0.5814
v	f	of	0.0001	0.0001	0.0002	0.0003	0.0035	0.0059	0.0139	0.0227
v	v	wives	0.0117	0.0110	0.0120	0.0073	0.7163	0.6706	0.7361	0.6364
v	ve	twelve	0.0046	0.0053	0.0041	0.0039	0.2801	0.3235	0.2500	0.3409
w	none	someone	0.0003	0.0005	0.0011	0.0016	0.0266	0.0388	0.0781	0.0732
w	u	subsequently	0.0022	0.0023	0.0014	0.0018	0.2021	0.1860	0.0938	0.0854
w	w	wound	0.0070	0.0079	0.0100	0.0152	0.6436	0.6357	0.6875	0.7073
w	wh	worthwhile	0.0014	0.0017	0.0020	0.0029	0.1277	0.1395	0.1406	0.1341
j	ea	beauty	0.0001	0.0002	0.0002	0.0003	0.0138	0.0225	0.0244	0.0526
j	ew	newspaper	0.0003	0.0005	0.0007	0.0013	0.0414	0.0562	0.0732	0.2632
j	i	view	0.0002	0.0003	0.0005	-	0.0207	0.0337	0.0488	-
j	u	vocabulary	0.0050	0.0047	0.0048	0.0008	0.5931	0.5506	0.5122	0.1579
j	ue	value	0.0005	0.0005	0.0005	-	0.0552	0.0562	0.0488	-
j	u-e	volume	0.0014	0.0012	0.0007	0.0005	0.1655	0.1348	0.0732	0.1053
j	y	youth	0.0009	0.0012	0.0020	0.0021	0.1103	0.1461	0.2195	0.4211

REFERENCES

Balota, D. A., Cortese M. J., Sergent-Marshall, S. D., Spieler, D. H. & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General,* **133***,* 283-316.

Baron, J., & Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance,* **2***,* 386-93.

Berndt, R.T., Reggia, J.A. & Mitchum, C.C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behaviour Research Methods, Instruments, & Computers*, **19**, 1-9.

Carney, E. (1994) *A Survey of English Spelling*. London: Routledge.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: dual-route and parallel-distributed-processing approaches. *Psychological Review,* **100***,* 589-608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001) DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, **108**, 204-256.

Cossu, G., Gugliotta, M., & Marshall, J. (1995). Acquisition of reading and written spelling in a transparent orthography: Two non-parallel processes? *Reading and Writing: An Interdisciplinary Journal*, **7***,* 9-22.

Cossu, G., Shankweiler, D., Liberman, I. Y., Katz, L., & Tola, G. (1988). Awareness of phonological segments and reading ability in Italian children. *Applied Psycholinguistics*, **9**, 1-16.

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., Van Daal, V. H. P., Polyzoe, N., Tsipa, N-M., & Petalas, M. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly*, **39**, 438-468.

Frith, U., Wimmer, H. & Landerl, K. (1998). Differences in Phonological Recoding in German- and English-Speaking Children. *Scientific Studies of Reading*, **2**, 31-54.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, **13**, 104-115.

Fry, E. (2004). Phonics: A large phoneme-grapheme frequency count revised. *Journal of Literacy Research*, **36**, 85-98.

Gontijo, P. F. D., Gontijo, I. & Shillcock, R. (2003). Grapheme-phoneme probabilities in British English. *Behavior Research Methods, Instruments & Computers*, **35**, 136-157.

Goswami, U., Porpodas, C., & Wheelwright, S. (1997). Children's orthographic representations in English and Greek. *European Journal of Psychology of Education*, **7**, 273-290.

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

Hino, Y., & Lupker, S. J. (2000). Effects of word frequency and spelling-to-sound regularity in naming with and without preceding lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, **26**, 166-183.

Hofland, K. & Johansson, S. (1982). *Word Frequencies in British and American English.* Bergen: The Norwegian Computing Centre for the Humanities.

Jared, D. (2002) Spelling–sound consistency and regularity effects in word naming. *Journal of Memory and Language*, **46**, 723–750.

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). Amsterdam: Elsevier Science Publishers.

Kessler, B., Treiman, R., & Mullennix, J. (2007). Feedback consistency effects in single-word reading. In E. L. Grigorenko & A. J. Naples (Eds.), *Single-word reading: Behavioral and biological perspectives* (pp. 159-174). Mahwah, NJ: Erlbaum.

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition*, **63**, 315–334.

Masterson, J., Stuart, M., Dixon, M. & Lovejoy, S. (2003). *Children's Printed Word Database. Economic and Social Research Council project (R00023406).* Retrieved September 28, 2007, Essex University Web site: http://www.essex.ac.uk/psychology/cpwd/

McGuinness, D. (1998). Why Children Can't Read: And What We Can Do About It. London: Penguin.

Norris, D., McQueen, J.M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, **23**, 299-370.

Oney, B. & Goldman, S.R. (1984). Decoding and comprehension skills in Turkish and English: Effects of regularity of grapheme-phoneme correspondences. *Journal of Educational Psychology*, **76**, 557-568.

Oxford University Press (2006). *The Oxford English Dictionary On-line*. Oxford : Oxford University Press. Retrieved June 1, 2006, from http://dictionary.oed.com/

Paulesu, E., Demonet, J.-F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., Cappa, S. F., Cossu, G., Habib, M., Frith, C. D., & Frith, U. (2001). Dyslexia: Cultural Diversity and Biological Unity, *Science*, **291**, 2165-2167.

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S. F., Cotelli, M., Cossu, G., Corte, F., Lorusso, M., Pesenti, S., Gallagher, A., Perani, D., Price, C., Frith, C. D., & Frith, U. (2000). A cultural effect on brain function. *Nature Neuroscience*, **3**, 91-96.

Perry C., Ziegler, J. C. & Coltheart, M. (2002). How predictable is spelling? Developing and testing metrics of phoneme-grapheme contingency. *The Quarterly Journal of Experimental Psychology*, **55A**, 897-915.

Perry C., Ziegler, J. C. & Zorzi, M. (2007). Nested incremental modelling in the development od computational theories: The CDP+ model of reading aloud. *Psychological Review*, **114**, 273-315.

Plaut, D. C. (1999). A connectionist approach to word reading and acquired dyslexia: extension to sequential processing. *Cognitive Science*, **23**, 543-568.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S. & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, **103**, 56-115.

Rastle, K. & Coltheart, M. (1999). Serial and strategic effects in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, **25**, 482-503.

Reggia, J. A., Marsland, P. M., & Berndt, R. S. (1988). Competitive dynamics in a dual-route connectionist model of print-to-sound transformation. *Complex Systems*, **2**, 509-547.

Scheerer, E. (1986). Orthographic and lexical access. In G. August (Ed.), *New Trends in Graphemics and Orthography* (pp. 262-286). Berlin: De Gruyter.

Seidenberg, M. S. & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**, 523-568.

Seymour, P. H. K., Aro M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, **94**, 143-174.

Spencer, K. A. (1999). Predicting word-spelling difficulty in 7- to 11-year-olds. *Journal of Research in Reading*, **22**, 283-292.

Spencer, K. A. (2001). Differential effects of orthographic transparency on dyslexia: word reading difficulty for common English words. *Dyslexia*, **7**, 217–228.

Spencer, K. A. (2007). Predicting children's word-spelling difficulty for common English words from measures of orthographic transparency, phonemic and graphemic length and word frequency. *British Journal of Psychology*, **98**, 305-338.

Spencer, K. A. (2008). *Predicting children's word-reading for common English words: the effect of word transparency and complexity.* Manuscript submitted for publication.

Spencer, L. H. & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology*, **94**, 1-28.

Stanovich, K. E., & Bauer, D. W. (1978). Experiments on the spelling-to-sound regularity effect in word recognition. *Memory & Cognition*, **6**, 410-415.

Stone, G. O., Vanhoy, M. & Van Orden, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology in visual word recognition. *Journal of Memory and Language*, **36**, 337-359.

Treiman, R., Mullennix, J., 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. (1967). English orthography: its graphical structure and its relation to sound. *Reading Research Quarterly*, **2**, 75-106.

Venezky, R. L. (1970). *The structure of English orthography.* The Hague: Moulton.

Waters, G. S., & Seidenberg, M. S. (1985). Spelling-sound effects in reading: time-course and decision criteria. *Memory & Cognition*, **13**, 557-572.

Ziegler, J. C., Jacobs, A. M., & Stone, G. O. (1996). Statistical analysis of the bidirectional inconsistency of spelling and sound in French. *Behavior Research Methods, Instruments and Computers*, **28**, 504-515.

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 and Computers*, **29**, 600-618.

Table 1

Summary of 5 metrics for the sonograph /eɪ/ ↔ <eigh>

	Frequency	Total	Probability
Sonograph	48	108571	0.0004
Grapheme	51	108571	0.0005
Phoneme	2248	108571	0.0207
Phoneme-grapheme PGC	48	2248	0.0214
Grapheme-phoneme GPC	48	51	0.9412

Table 2

Correlations among five word metrics for Corpus A and Corpus C

	Grapheme probability				Grapheme-phoneme correspondence			
	1	2	3	4	5	6	7	8
1. Corpus A: G 3K	-	1.00 **	.98 **	.91 **	-.60 **	-.58 **	-.62 **	-.53 **
2. Corpus A: G 2K		-	.99 **	.92 **	-.57 **	-.59 **	-.63 **	-.56 **
3. Corpus A: G 1K			-	.93 **	-.57 **	-.60 **	-.64 **	-.57 **
4. Corpus C: G 1K				-	-.54 **	-.56 **	-.58 **	-.58 **
5. Corpus A: GP 3K					-	.98 **	.96 **	.93 **
6. Corpus A: GP 2K						-	.98 **	.93 **
7. Corpus A: GP 1K							-	.95 **
8. Corpus C: GP 1K								-
9. Corpus A: P 3K								
10. Corpus A: P 2K								
11. Corpus A: P 1K								
12. Corpus C: P 1K								
13. Corpus A: SG 3K								
14. Corpus A: SG 2K								
15. Corpus A: SG 1K								
16. Corpus C: SG 1K								
17. Corpus A: PG 3K								
18. Corpus A: PG 2K								
19. Corpus A: PG 1K								
20. Corpus C: PG 1K								

Note. G = Grapheme probability of occurrence; P = Phoneme probability of occurrence; GP probability = Grapheme-phoneme correspondence (reading) probability; PG = Phoneme-grapheme correspondence (spelling) probability; SG = Sonograph probability of occurrence.

Corpus A is derived from Hofland, K. & Johansson, S. (1982): 3K = 3,220 words; 2K = 2,019 words; 1K = 928 words.

Corpus C is derived from Masterson, J., Stuart, M., Dixon, M. & Lovejoy, S. (2003): 1K = 971 words.

* p < .05. ** p < .01.

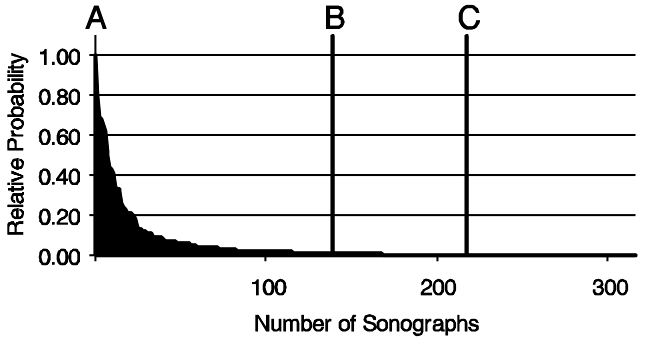
Table 2

(Continued)

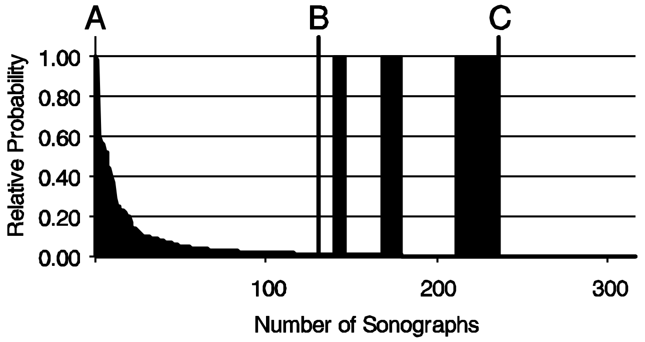
Phoneme probability				Sonograph probability				Phoneme-grapheme correspondence			
9	10	11	12	13	14	15	16	17	18	19	20
.03	.07	.08	.12	.60 **	.64 **	.63 **	.51 **	.55 **	.56 **	.53 **	.44 **
.06	.05	.06	.12	.61 **	.64 **	.63 **	.51 **	.55 **	.57 **	.55 **	.43 **
.07	.06	.05	.10	.58 **	.61 **	.64 **	.52 **	.53 **	.55 **	.56 **	.46 **
.10	.10	.09	.11	.56 **	.57 **	.60 **	.58 **	.50 **	.50 **	.51 **	.50 **
.10	.12	.16 *	.16 *	.17 **	.15 *	.12	.17 *	.09	.08	.04	.07
.11	.11	.16 *	.13	.15 *	.13 *	.10	.16 *	.09	.06	.02	.07
.12	.12	.13 *	.13	.11	.08	.06	.13	.04	.02	.00	.05
.06	.04	.08	.11	.12	.11	.07	.19 **	.09	.08	.05	.11
-	1.00 **	.98 **	.92 **	.09	.11	.14 *	.10	-.41 **	-.41 **	-.42 **	-.44 **
	-	.98 **	.92 **	.12	.11	.14 *	.09	-.39 **	-.41 **	-.42 **	-.45 **
		-	.94 **	.17 *	.15 *	.14 *	.14	-.35 **	-.38 **	-.42 **	-.42 **
			-	.19 **	.17 *	.19 **	.19 **	-.29 **	-.32 **	-.34 **	-.39 **
				-	.98 **	.93 **	.79 **	.85 **	.82 **	.74 **	.64 **
					-	.96 **	.81 **	.84 **	.84 **	.76 **	.66 **
						-	.82 **	.80 **	.80 **	.80 **	.66 **
							-	.71 **	.71 **	.68 **	.77 **
								-	.99 **	.95 **	.88 **
									-	.97 **	.90 **
										-	.90 **
											-

Figure 1. Sonograph probability profiles for Corpus C and Corpus A. For comparative purposes relative probability values are expressed as a proportion of the highest frequency sonograph. Position B indicates relative probabilities < .01. Position C = total number of sonographs for the corpus. Bars in area BC indicate new sonographs for the corpus compared with the smaller corpus above it.

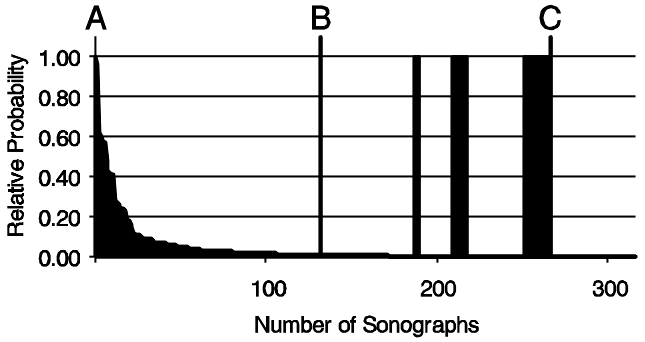
Corpus C 1K



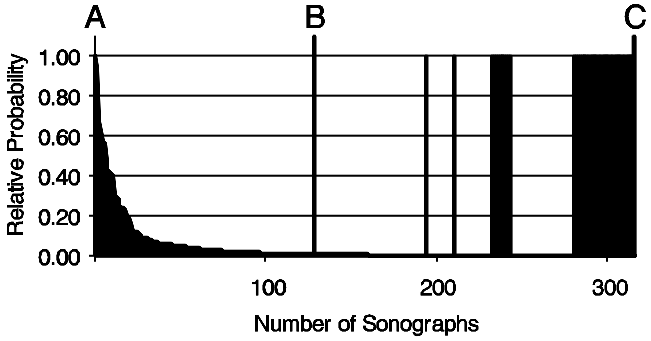
Corpus A 1K



Corpus A 2K



Corpus A 3K



[1] In this paper only the terms transparent (rather than shallow) and opaque (rather than deep) will be used.

[2] Within British English there is some variation from region to region, and the metrics described should be applied with caution outside southern Britain.

[3] For comparative purposes probabilities are expressed relative to the most frequent sonograph.