

The regularity of orthographies varies, with those that are regular classed as “transparent”, and those with irregular correspondences classed as “opaque”. Finnish and Turkish are symmetrically transparent for both reading and spelling, whereas Greek and German are asymmetric, being less transparent for spelling than reading, with letters that have clearly defined pronunciations, but phonemes that have alternative spellings. It has been proposed (Katz & Frost, 1992) that the variation between orthographies leads to differences in processing and, consequently, transparent orthographies make literacy skills easier to learn (Oney & Goldman, 1984; Spencer & Hanley, 2003). Accuracy levels are lower and reading speed is slower for opaque languages for both normal and dyslexic children (Cossu, Gugliotta, & Marshall, 1995; Ellis et al., 2004; Frith, Wimmer, & Landerl, 1998; Landerl, Wimmer, & Frith, 1997).

Detailed linguistic analyses have measured the orthographic body/phonological rime (e.g. as in *seen*, the orthographic body <een> has the phonological rime /i:n/) transparency of English (Ziegler, Stone, & Jacobs, 1997), and French (Ziegler, Jacobs, & Stone, 1996) for both reading and spelling, and allow comparisons of relative transparency. For example, 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 which the ease of pronunciation depends on the relative orthographic body transparency of a word. Words having body letter patterns with the same phonological rime that are always pronounced in the same way are classed as consistent. The less typical the pronunciation, the greater the word reading difficulty, for words classed as inconsistent. Treiman, Mullennix, Bijeljac-Babic, and Richmond-Welty (1995) see

dichotomous classification (consistent/inconsistent) as inadequate, and claim that only continuous representations of consistency allow detailed examination of transparency effects, and Plaut (1999) suggests that the language mechanism gradually picks up on this continuous statistical structure among written and spoken words.

Languages may also be classified at the fine-grain grapheme-phoneme level for reading and spelling. English has been studied extensively at this level (Hanna, Hanna, & Hodges, 1966; Venezky, 1967; Berndt, Reggia, & Mitchum, 1987; Carney, 1994; Gontijo, Gontijo, & Shillcock, 2003). The present study applies this approach to the Greek language.

Variations of fine-grain word transparency for English formed an integral part of the early serial dual route models of skilled reading (Coltheart, Curtis, Atkins, & Haller, 1993). Words were divided into graphemes and their associated phonemes, which were termed grapheme-phoneme correspondences (GPC) when associated with reading, and phoneme-grapheme correspondences (PGC) when associated with spelling. Separate GPC and PGC conditional probabilities could be calculated from the same sound-letter components. More recently it has been conceded that algorithm-derived GPC rules do not work for the DRC model, which now features hard-wired GPC rules (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) and with this there is an explicit admission that the DRC model does not offer any account of how reading is learned (Coltheart, 2006). This has led to the development of the connectionist dual process (CDP) model, which is claimed to be superior because it is highly sensitive to the graded statistical consistency of spelling-sound relationships at multiple grain sizes (from letters to word bodies) and has a stronger developmental strand than alternative computer models (Perry, Ziegler and Zorzi, 2007).

Different sources and corpus sizes have been used when calculating word metrics. Large poly-syllabic adult corpora of 20,000 and 17,310 words were used by Venezky (1967) and Hanna et al. (1966), whereas Seidenberg and McClelland (1989) and Coltheart et al. (1993) used a corpus of only 2,897 monosyllabic words. McGuinness (1998) argued that spelling patterns for smaller corpora of more common words would differ from more extensive analyses, but small corpora have been shown to have substantially the same fine-grain structure as larger corpora for English (Spencer, 2009). This offers researchers the opportunity to prepare metrics that are closely associated with experimental conditions, such as metrics derived from specific children's texts, rather than general adult corpora, for experiments with children as subjects. Spencer (2007) demonstrated that in multiple regression analyses, frequency metrics derived from such a children's corpus (Children's Printed Word Database; Masterson, Stuart, Dixon and Lovejoy, 2003) predicted substantially more individual variance in a spelling task, for school years 2 – 6, than values derived from larger adult corpora. However, smaller corpora do under-estimate the number of grapheme-phoneme alignments (sonographs, see below), although the under-estimated sonographs fall in the long tail of very low frequencies (probabilities < 0.01) for English.

Before reading or spelling conditional probabilities can be 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. It is this type frequency that is used to calculate both GPC or PGC probability values. Spencer (2009) used the term *sonograph* to describe this directionless item, and demonstrated how five basic metrics for the fine-grain level of analysis may be calculated from it. This approach has been adopted for the present analysis of the

Greek language. Table 1 illustrates the calculation of the five metrics for one sonograph: /i/-<ει>. The 4,162 words in the present corpus are made up of a total of 27,585 constituent phonemes and their associated graphemes (27,585 sonographs). Of these there are 2,412 occurrences of the phoneme /i/, which is represented by 6 graphemes (i.e. there are 6 sonographs associated with the /i/ phoneme). The <ει> grapheme occurs 437 times in association with /i/ (the remaining 1,975 occurrences are distributed among the other 5 graphemes associated with /i/). Thus, the unique /i/-<ει> sonograph occurs 437 times in the total number of 27,585 sonographs, with a probability of occurrence of 0.0158. The grapheme <ει> is usually associated with the /i/ phoneme (437 occurrences), but in a small number of cases (10) it is associated with other phonemes. So, although it is highly predictable, its probability (.98) falls short of unity.

Table 1 about here

Greek orthography

Hatzigeorgiou Mikros and Karayiannis (2001) carried out analyses of the characteristics of written Greek, based on the Hellenic National Corpus (HNC), a corpus of written Modern Greek developed by the Institute for Language and Speech Processing. They reported that the average word length is 5.45 letters for the 13 million words in their corpus, and that the most familiar words are shorter than the average length of the whole corpus. Ktori, van Heuven and Pitchford (2008) recently developed a lexical database of Modern Greek based on the Lexicon of Common Modern Greek and HNC. Their analyses revealed that the majority of Greek words are between eight- and nine-letters in length, with a mean word length of 9.07 letters,

but about 50% of the summed word frequencies were accounted for by words with four or fewer letters, leading to the conclusion that the distribution of word length in Greek is compatible with Zipf's 'principle of least effort' (Zipf, 1949). The average word in terms of frequency was found to be 5.7 letters long, comparable to that reported for HNC (Hatzigeorgiu et al., 2001; Mikros, Hatzigeorgiu, & Carayannis, 2005).

For broad phonetic transcription, modern Greek has 32 phonemes, of which 5 are vowels. These are written with 25 letters (including one end-only form), of which 7 correspond to vowels in isolation. Protopapas and Vlahou, (in press) suggest that most words can be read correctly on the basis of the letter sequence alone, without the need for morphological or lexical information, and have estimated Greek to have an overall feedforward consistency of 96% at the grapheme-phoneme level. This level of reading consistency has led Porpodas (2006) and Seymour, Aro, and Erskine (2003) to classify Greek as a shallow orthography. However, Porpodas also recognizes that for spelling, Greek is phonologically opaque, with one-to-many mappings, which means that spelling is not always predictable. Protopapas and Vlahou estimate the feedback token consistency to be 83%.

Thus, for reading correctly (but not necessarily fluently, e.g., Porpodas, 1999; Loizidou-Ieridou, 2007) phonological decoding is sufficient. On the other hand, very few words are regular in terms of one-to-one sound-letter correspondences, and the irregularities (mostly involving the representation of three of the five vowels) are usually affected by morphology, and reflect semantic and grammatical distinctions. For example, the decision between *o* and *ω* for the /o/ sound can be made using the rule that verbs are spelled with <*ω*> (e.g., *βάζω*: to put) and neutral nouns <*o*> (e.g., *βάζο*: flower pot). Thus, correct spelling depends partly on decoding strategies and

partly on morphological and syntactic knowledge (e.g., Chliounaki & Bryant, 2002; Porpodas, 1999; Loizidou-Ieridou, Masterson & Hanley, 2009).

For Modern Greek the stress-accent coincides with the written accent (´) of a word, e.g., μήλο - /'milo/. Only vowel letters bear a punctuation mark, additionally the accent mark is only used when the word is of two or more syllables and when the word is written in lowercase (capital letters do not bear diacritics). There is also the diacritic dieresis (¨) to disambiguate single vowel graphemes from digraphs, e.g. αῖ is pronounced as two separate vowels as in αἰβάλη - /aiva'li/.

METHOD

Corpora and alignments

For the present analysis a database was generated from the language text books used for Greek language instruction in the first two Primary school grades in Cyprus (Karantzola, Kurdi, Spanelli, & Tsiagakani, 2008). This process resulted in a sample of 4,162 poly-syllabic words.

To aid comparability with recently published sources of information concerning the Greek language (Hellenic National Corpus [HNC]; Hatzigeorgiu et al., 2000; <http://hnc.ilsp.gr>), phonemic definitions and alignments for the children's corpus were obtained from the Psycho-linguistic Resources available at the Institute for Language and Speech Processing (<http://speech.ilsp.gr/iplr/>). The analysis by Protopapas and Vlahou (in press) used 217,644 unique word forms (types) accounting for 29,557,090 occurrences (tokens), based on 36 letters:

Of the 24 letters in the Greek alphabet, seven (the “vowel letters”) have variants bearing diacritics. Specifically, all 7 may be accompanied by an acute accent, indicating stress. Two of these may carry diaeresis,

indicating exception from digraph combinations. Because both types of diacritics are useful in phonological or lexical disambiguation, and because they are dictated by current spelling rules and their omission is always a spelling error, the variants of these letters with diacritics (stress mark only, diaeresis only, or both) were retained in the counts as separate letters. Including the word-final variant ζ , a total of 36 letters were used in the analyses. (Protopapas & Vlahou, in press)

Their alignments are based on 37 phonemes:

The resulting set of 32 phonemes, 5 of which are vowels, suffice to accurately and completely represent phonetically (broadly, at the surface realization) every Greek word in standard modern pronunciation typical of major cities such as Athens. To retain stress information, aiding in disambiguation, stressed vowels were represented as separate phonemes, bringing the total number of phonemes to 37. (Protopapas & Vlahou, in press)

The alignment process for the full corpus of 4,162 words resulted in a total of 100 sonographs, whereas the much larger corpus (217,664 words) of Protopapas and Vlahou (in press) yielded 118. In addition to the full corpus, a series of smaller corpora was created in order to explore the sonograph profiles for corpora of varying sizes. Spencer (2009) demonstrated that for English, increasing corpus size does not substantially alter the orthographic profile, it simply introduces additional low probability sonographs. The number of words and sonographs for the four corpora are:

Corpus .21, number of words (nw) = 211, number of sonographs (nsg) = 63, frequency range (fr) ≥ 10 ; Corpus .71, nw = 706, nsg = 83, fr ≥ 3 ; Corpus 1.66, nw = 1660, nsg = 93, fr > 1 ; Corpus 4.16, nw = 4162, nsg = 100, fr = 1.

Frequency counts for each sonograph for the four corpora were produced, from which reading and spelling probabilities were calculated. The following measures were calculated: phoneme and grapheme frequency and probability of occurrence, sonograph probability of occurrence, grapheme-phoneme correspondence (reading) probability and phoneme-grapheme correspondence (spelling) probability.

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: .98 for grapheme, and .96 for phoneme probabilities of occurrence; .96 for grapheme-phoneme (reading) correspondence probabilities, .96 for phoneme-grapheme (spelling) correspondence probabilities; and .97 for sonograph probabilities of occurrence. This suggests that the metrics describing Greek remain substantially the same for corpora of varying size, from sources of writing for children. The present analysis finds that the core probability values for Greek may be obtained from relatively small samples of words, and this is reflected in Figure 1. For the most frequent words in the children's .21K corpus of 63 sonographs, all are in the relative probability¹ of occurrence range .01 to 1.0 (shown between A and B, Figure 1). The larger corpus .70K corpus has 20 additional sonographs, most of which are of very low probability (< 0.01), indicated by the bars between B and C. This pattern is repeated for the increasingly larger corpora. The difference between corpus size of .70K and 1.66K is an additional 10

very low frequency sonographs (the bars between B and C); and between 1.66K and 4.10K, there is an additional 7 very low frequency sonographs.

The purpose of calculating descriptive metrics for a transparent language such as Greek is to provide continuous measures of the variables that may support current linguistic data (Hellenic National Corpus [HNC]; Hatzigeorgiu et al., 2000; Ktori, van Heuven, & Pitchford, 2008; Protopapas & Vlahou, in press) which may be applied in linguistic and psychological studies. This has usually involved the decomposition of very large corpora, but it appears that this study confirms Spencer's (2009) results for English, which demonstrated that smaller samples of words produced metrics that are very similar to those from larger samples.

Table 2 also demonstrates a relationship that reflects the difference between reading and spelling in Greek. The negative correlations between phoneme probability of occurrence and phoneme-grapheme correspondence probability, although small, reflect the nature of the spelling metric: the more frequent phonemes tend to be vowels, which are associated with more than one grapheme and consequently have PG values less than 1.0; less frequent phonemes, mainly consonants, tend to have PG probabilities close to 1.0 because they are often associated with only one grapheme.

Figure 2 compares the sonograph probabilities for Greek with those for UK English (based on Spencer, 2009). The long tail observed in the opaque English orthography is not present in the relatively transparent Greek distribution. This clearly suggests that literacy skills for Greek will be easier to learn because there are fewer sonographs, and a larger proportion of these are more frequent and provide more opportunities for learning. This is reflected in the additional adult processing resources observed by Paulesu et al. (2000) when comparing PET (Positron Emission

Tomography) brain scan activity for a transparent (Italian) and two opaque (English and French) languages.

However, Figure 3 shows Zipfian plots for words and phonemes in both languages, demonstrating their similarity at this structural level. Zipf's distribution of the logarithm of the rank of a signal (word, phoneme) against the logarithm of the frequency of occurrence for a human language (Zipf 1936, 1949) provides a function of its potential capacity for communication. Zipf's law is based on the 'Principle of Least Effort' in which human speech and language are structured optimally by two opposing forces, unification and diversification. This process results in a balance that can be statistically represented by a regression coefficient (or slope) of -1.00 , regressing the log of the rank on the log of actual frequency of occurrence. This applies to a multitude of diverse human languages, including modern Greek.

Hatzigeorgiu et al. (2001) found the slope for modern Greek words to be -0.98 , which compares favourably with the present study's value of -0.97 . As would be expected for a relatively transparent language, the regression slope for Greek sonographs is also close to -1.0 . For English, reflecting its lack of orthographic transparency, the slope moves away from the ideal, and in Figure 3, covering cumulative frequencies up to 99% (198 of the 316 sonographs) in order to avoid distortion by very low frequency sonographs (see Martindale, Gusein-Zade, McKenzie & Borodovsky, 1996), it has a value of -1.4 . The regression coefficient for the full range of English sonographs is -1.8 . Clearly the balancing processes identified by Zipf do not appear to be influencing English orthography, but generally do apply to Greek. However, in a totally transparent orthography the number of phonemes would dictate the number of sonographs, and this is not the case for the present Greek corpus, which has 37 phonemes and 100 sonographs. Appendix A demonstrates that, although there are

more sonographs than phonemes, reading probabilities are generally high, with the mean dominant reading probability for vowels being 0.95. This is not the case for spelling probabilities. Appendix B shows that most vowels have spelling probabilities that are less than unity, and that the mean dominant spelling probability for vowels is 0.79, confirming claims that Greek is less transparent for spelling than reading. These continuous spelling correspondence values should influence Greek children's spelling performance, in a similar manner to the influence of English phoneme-grapheme correspondences on spelling (Spencer, 2007).

TABLE 2 about here

FIGURES 1, 2 & 3 about here

CONCLUSION

Word metrics have tended to focus on feed-forward and feed-backward metrics to account for variations in reading and spelling performance. The present paper includes the primary fine-grain metrics (sonograph probabilities), from which these secondary metrics are calculated. They 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 the transparently asymmetric Greek language.

APPENDIX A: Grapheme and GPC Probability

Grapheme probability = Grapheme probability of occurrence.

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

Grapheme	Grapheme probability				Phoneme	GPC probability			
	Corpus Size					Corpus Size			
	4.10 K	1.66 K	0.70 K	0.21 K		4.10 K	1.66 K	0.70 K	0.21 K
α	0.0972	0.0941	0.0917	0.0895	a	0.9948	0.9891	0.9758	0.9054
					'a	0.0052	0.0109	0.0242	0.0946
ά	0.0389	0.0457	0.0540	0.0484	'a	1.0000	1.0000	1.0000	1.0000
αι	0.0055	0.0058	0.0058	0.008	ε	0.9934	0.9821	0.9524	0.857
					'e	0.0066	0.0179	0.0476	0.1429
αί	0.0023	0.0029	0.0025	0.0012	'e	1.0000	1.0000	1.0000	1.0000
β	0.0116	0.0125	0.0114	0.0036	v	1.0000	1.0000	1.0000	1.0000
γ	0.0199	0.0215	0.0199	0.0157	ɣ	0.6095	0.6029	0.5833	0.6923
					ɰ	0.3832	0.3971	0.417	0.308
					ŋ	0.0073	-	-	-
γγ	0.0005	0.0003	0.0003	-	ʝ	0.7333	0.3333	-	-
					g	0.2667	0.6667	1.0000	-
γει	0.0000	0.0001	0.0003	-	ɰ	1.0000	1.0000	1.0000	-
γι	0.0015	0.0022	0.0030	0.0048	ɰ	1.0000	1.0000	1.0000	1.0000
γκ	0.0014	0.0009	0.0008	-	g	0.7436	0.5556	-	-
					ʝ	0.2564	0.4444	1.0000	-
γκι	0.0001	0.0001	-	-	ʝ	1.0000	1.0000	-	-
γυ	0.0001	0.0001	-	-	ɰ	1.0000	1.0000	-	-
δ	0.0177	0.0192	0.0216	0.019	ð	1.0000	1.0000	1.0000	1
ε	0.0580	0.0561	0.0512	0.0496	ε	0.9675	0.9707	0.9676	0.9268
					'e	0.0325	0.0293	0.0324	0.0732
έ	0.0235	0.0265	0.0277	0.0302	'e	1.0000	1.0000	1.0000	1.0000
ει	0.0162	0.0179	0.0199	0.0133	i	0.9776	0.9598	0.9167	0.9091
					'i	0.0134	0.0230	0.0556	0.0909
					ɰ	0.0045	0.0115	0.0139	-
					ç	0.0045	0.0057	0.0139	-
εί	0.0058	0.0063	0.0066	0.0133	'i	1.0000	1.0000	1.0000	1.0000
ζ	0.0083	0.0064	0.0047	0.0024	z	1.0000	1.0000	1.0000	1.0000

APPENDIX A: (Continued)

Grapheme	Grapheme probability				Phoneme	GPC probability			
	Corpus Size					Corpus Size			
	4.10 K	1.66 K	0.70 K	0.21 K		4.10 K	1.66 K	0.70 K	0.21 K
η	0.0249	0.0246	0.019	0.027	i	0.9869	0.9791	0.956	0.909
					ʰi	0.0102	0.0209	0.0441	0.0909
					ɰ	0.0029	-	-	-
ή	0.0126	0.0148	0.0163	0.0169	ʰi	1.0000	1.0000	1.0000	1.0000
θ	0.0108	0.0097	0.0072	0.0097	θ	1.0000	1.0000	1.0000	1.0000
ι	0.0371	0.0369	0.0352	0.0363	i	0.8074	0.7827	0.7402	0.7667
					ɰ	0.1173	0.1142	0.1417	0.1000
					ç	0.0674	0.0836	0.0866	0.0667
					ɲ	0.0059	0.0139	0.0157	0.0333
					ʰi	0.0020	0.0056	0.0157	0.0333
ί	0.0158	0.0171	0.0180	0.0169	ʰi	1.0000	1.0000	1.0000	1.0000
ϊ	0.0006	0.0004	-	-	i	1.0000	1.0000	-	-
κ	0.0412	0.0408	0.0380	0.0351	k	0.6989	0.7053	0.7445	0.7586
					c	0.3011	0.2947	0.2555	0.2414
κι	0.0018	0.0013	0.0028	0.0012	c	1.0000	1.0000	1.0000	1.0000
κκ	0.0004	0.0004	0.0006	0.0012	c	0.9000	1.0000	1.0000	1.0000
					k	0.1000	-	-	-
λ	0.0331	0.0324	0.0310	0.0302	l	1.0000	1.0000	1.0000	1.0000
λει	0.0003	0.0004	0.0003	-	λ	1.0000	1.0000	1.0000	-
λι	0.0021	0.0020	0.0022	0.0012	λ	1.0000	1.0000	1.0000	1.0000
λλ	0.0024	0.0029	0.0039	0.0073	l	1.0000	1.0000	1.0000	1.0000
λλι	0.0001	0.0002	-	-	λ	1.0000	1.0000	-	-
μ	0.0332	0.0324	0.0332	0.0387	m	1.0000	1.0000	1.0000	1.0000
μμ	0.0008	0.0007	0.0008	0.0012	m	1.0000	1.0000	1.0000	1.0000
μπ	0.0041	0.0046	0.0047	0.0048	b	1.0000	1.0000	1.0000	1.0000
ν	0.0495	0.0471	0.0396	0.0423	n	1.0000	1.0000	1.0000	1.0000
νι	0.0016	0.0012	0.0008	-	ɲ	1.0000	1.0000	1.0000	-
νν	0.0004	0.0007	0.0006	-	n	1.0000	1.0000	1.0000	-
ννι	0.0001	0.0001	0.0003	-	ɲ	1.0000	1.0000	1.0000	-

APPENDIX A: (Continued)

Grapheme	Grapheme probability				Phoneme	GPC probability			
	Corpus Size					Corpus Size			
	4.10 K	1.66 K	0.70 K	0.21 K		4.10 K	1.66 K	0.70 K	0.21 K
voi	0.0001	-	-	-	ɲ	1.0000	-	-	-
vτ	0.0080	0.0060	0.0044	0.0024	d	1.0000	1.0000	1.0000	1.0000
vτζ	0.0002	0.0002	0.0003	-	ɖ	1.0000	1.0000	1.0000	-
ξ	0.0086	0.0077	0.0086	0.0048	ks	1.0000	1.0000	1.0000	1.0000
ο	0.0410	0.0396	0.0404	0.0484	o	0.9991	0.9974	0.9932	1.0000
					'o	0.0009	0.0026	0.0068	-
ό	0.0189	0.0210	0.0247	0.0326	'o	1.0000	1.0000	1.0000	1.0000
οι	0.0060	0.0059	0.0086	0.007	i	0.8424	0.7368	0.6129	0.5
					ç	0.1333	0.2632	0.3871	0.5000
					ɲ	0.0242	-	-	-
οί	0.0014	0.0009	0.0011	-	'i	1.0000	1.0000	1.0000	-
ου	0.0156	0.0146	0.0152	0.0109	u	0.9930	0.9930	0.9818	1.0000
					'u	0.0070	0.0070	0.0182	-
ού	0.0123	0.0114	0.0078	0.0048	'u	1.0000	1.0000	1.0000	1.0000
π	0.0350	0.0387	0.0452	0.0508	p	1.0000	1.0000	1.0000	1.0000
ππ	0.0003	0.0005	0.0006	-	p	1.0000	1.0000	1.0000	-
ρ	0.0577	0.0580	0.0615	0.0532	r	1.0000	1.0000	1.0000	1.0000
ρρ	0.0003	0.0002	-	-	r	1.0000	1.0000	-	-
σ	0.0504	0.0442	0.0391	0.0399	s	0.9511	0.9721	0.9929	1.0000
					z	0.0489	0.0279	0.0071	-
ς	0.0287	0.0307	0.0330	0.0375	s	1.0000	1.0000	1.0000	1.0000
σσ	0.0008	0.0011	0.0011	-	s	1.0000	1.0000	1.0000	-
τ	0.0499	0.0486	0.0485	0.0665	t	1.0000	1.0000	1.0000	1.0000
τζ	0.0003	0.0003	-	-	ɖ	1.0000	1.0000	-	-
τσ	0.0018	0.0015	0.0019	0.0012	ts	1.0000	1.0000	1.0000	1.0000

APPENDIX A: (Continued)

Grapheme	Grapheme probability				Phoneme	GPC probability			
	Corpus Size					Corpus Size			
	4.10 K	1.66 K	0.70 K	0.21 K		4.10 K	1.66 K	0.70 K	0.21 K
υ	0.0153	0.0145	0.0136	0.0097	i	0.7470	0.7092	0.7143	0.3750
					f	0.1371	0.2128	0.1837	0.5000
					v	0.1064	0.0638	0.0612	-
					j	0.0047	0.0071	0.0204	0.1250
					ɲ	0.0024	0.0071	0.0204	-
					ç	0.0024	-	-	-
ύ	0.0070	0.0065	0.0061	0.0085	'i	0.7098	0.7619	0.8636	1.0000
					v	0.1813	0.1587	0.1364	-
					f	0.1088	0.0794	-	-
υι	0.0000	-	-	-	ç	1.0000	-	-	-
φ	0.0162	0.0164	0.0155	0.0097	f	1.0000	1.0000	1.0000	1.0000
χ	0.0171	0.0184	0.0197	0.0121	x	0.7288	0.7039	0.7042	0.3000
					ç	0.2712	0.2961	0.2958	0.7000
χι	0.0004	0.0004	-	-	ç	1.0000	1.0000	-	-
ψ	0.0033	0.0031	0.0025	0.0012	ps	1.0000	1.0000	1.0000	1.0000
ω	0.0145	0.0139	0.0166	0.0157	o	0.9875	0.9704	0.9667	1.0000
					'o	0.0125	0.0296	0.0333	-
ώ	0.0075	0.0065	0.0083	0.0133	'o	1.0000	1.0000	1.0000	1.0000

APPENDIX B: Phoneme, PGC and Sonograph Probability

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

Ph	Gr	Phoneme probability				PGC probability				Sonograph probability			
		Corpus Size				Corpus Size				Corpus Size			
		4.10 K	1.66 K	0.70 K	0.21 K	4.10 K	1.66 K	0.70 K	0.21 K	4.10 K	1.66 K	0.70 K	0.21 K
a	α	0.0966	0.0930	0.0895	0.0810	1.0000	1.0000	1.0000	1.0000	0.0966	0.0930	0.0895	0.0810
'a	ά	0.0394	0.0468	0.0562	0.0568	0.9871	0.9780	0.9606	0.8511	0.0389	0.0457	0.0540	0.0484
'a	α					0.0129	0.0220	0.0394	0.1489	0.0005	0.0010	0.0022	0.0085
b	μπ	0.0041	0.0046	0.0047	0.0048	1.0000	1.0000	1.0000	1.0000	0.0041	0.0046	0.0047	0.0048
c	κ	0.0146	0.0138	0.0130	0.0109	0.8507	0.8731	0.7447	0.7778	0.0124	0.0120	0.0097	0.0085
c	κΙ					0.1269	0.0970	0.2128	0.1111	0.0018	0.0013	0.0028	0.0012
c	κκ					0.0224	0.0299	0.0426	0.1111	0.0003	0.0004	0.0006	0.0012
ç	χ	0.0084	0.0106	0.0125	0.0145	0.5494	0.5146	0.4667	0.5833	0.0046	0.0054	0.0058	0.0085
ç	Ι					0.2961	0.2913	0.2444	0.1667	0.0025	0.0031	0.0030	0.0024
ç	οΙ					0.0944	0.1456	0.2667	0.2500	0.0008	0.0015	0.0033	0.0036
ç	χΙ					0.0429	0.0388	-	-	0.0004	0.0004	-	-
ç	εΙ					0.0086	0.0097	0.0222	-	0.0001	0.0001	0.0003	-
ç	υ					0.0043	-	-	-	0.0000	-	-	-
ç	υΙ					0.0043	-	-	-	0.0000	-	-	-
d	ντ	0.0080	0.0060	0.0044	0.0024	1.0000	1.0000	1.0000	1.0000	0.0080	0.0060	0.0044	0.0024
ð	ð	0.0177	0.0192	0.0216	0.0193	1.0000	1.0000	1.0000	1.0000	0.0177	0.0192	0.0216	0.0193
ð	τζ	0.0005	0.0005	0.0003	-	0.6429	0.6000	-	-	0.0003	0.0003	-	-
ð	ντζ					0.3571	0.4000	1.0000	-	0.0002	0.0002	0.0003	-
f	φ	0.0191	0.0200	0.0180	0.0145	0.8498	0.8205	0.8615	0.6667	0.0162	0.0164	0.0155	0.0097
f	υ					0.1103	0.1538	0.1385	0.3333	0.0021	0.0031	0.0025	0.0048
f	ύ					0.0399	0.0256	-	-	0.0008	0.0005	-	-
g	γκ	0.0012	0.0007	0.0003	-	0.8788	0.7143	-	-	0.0011	0.0005	-	-
g	γγ					0.1212	0.2857	1.0000	-	0.0001	0.0002	0.0003	-
γ	γ	0.0121	0.0130	0.0116	0.0109	1.0000	1.0000	1.0000	1.0000	0.0121	0.0130	0.0116	0.0109

APPENDIX B: (Continued)

Ph	Gr	Phoneme probability				PGC probability				Sonograph probability			
		Corpus Size				Corpus Size				Corpus Size			
		4.10 K	1.66 K	0.70 K	0.21 K	4.10 K	1.66 K	0.70 K	0.21 K	4.10 K	1.66 K	0.70 K	0.21 K
i	ι	0.0874	0.0851	0.0773	0.0713	0.3425	0.3394	0.3369	0.3898	0.0299	0.0289	0.0260	0.0278
i	η					0.2807	0.2826	0.2330	0.3390	0.0245	0.0241	0.0180	0.0242
i	ει					0.1812	0.2017	0.2366	0.1695	0.0158	0.0172	0.0183	0.0121
i	υ					0.1310	0.1208	0.1254	0.0508	0.0115	0.0103	0.0097	0.0036
i	οι					0.0576	0.0507	0.0681	0.0508	0.0050	0.0043	0.0053	0.0036
i	ϊ					0.0070	0.0048	-	-	0.0006	0.0004	-	-
ʲ	ί	0.0411	0.0451	0.0499	0.0605	0.3850	0.3781	0.3611	0.2800	0.0158	0.0171	0.0180	0.0169
ʲ	ή					0.3057	0.3280	0.3278	0.2800	0.0126	0.0148	0.0163	0.0169
ʲ	εί					0.1410	0.1390	0.1333	0.2200	0.0058	0.0063	0.0066	0.0133
ʲ	ύ					0.1207	0.1093	0.1056	0.1400	0.0050	0.0049	0.0053	0.0085
ʲ	οί					0.0344	0.0205	0.0222	-	0.0014	0.0009	0.0011	-
ʲ	η					0.0062	0.0114	0.0167	0.0400	0.0003	0.0005	0.0008	0.0024
ʲ	ει					0.0053	0.0091	0.0222	0.0200	0.0002	0.0004	0.0011	0.0012
ʲ	ι					0.0018	0.0046	0.0111	0.0200	0.0001	0.0002	0.0006	0.0012
ʃ	γγ	0.0009	0.0006	0.0008	-	0.4583	0.1667	-	-	0.0004	0.0001	-	-
ʃ	γκ					0.4167	0.6667	1.0000	-	0.0004	0.0004	0.0008	-
ʃ	γκι					0.1250	0.1667	-	-	0.0001	0.0001	-	-
ʒ	γ	0.0138	0.0154	0.0172	0.0145	0.5526	0.5533	0.4839	0.3333	0.0076	0.0085	0.0083	0.0048
ʒ	ι					0.3158	0.2733	0.2903	0.2500	0.0044	0.0042	0.0050	0.0036
ʒ	γι					0.1079	0.1400	0.1774	0.3333	0.0015	0.0022	0.0030	0.0048
ʒ	γυ					0.0053	0.0067	-	-	0.0001	0.0001	-	-
ʒ	ει					0.0053	0.0133	0.0161	-	0.0001	0.0002	0.0003	-
ʒ	η					0.0053	-	-	-	0.0001	-	-	-
ʒ	υ					0.0053	0.0067	0.0161	0.0833	0.0001	0.0001	0.0003	0.0012
ʒ	γει					0.0026	0.0067	0.0161	-	0.0000	0.0001	0.0003	-
k	κ	0.0288	0.0288	0.0283	0.0266	0.9987	1.0000	1.0000	1.0000	0.0288	0.0288	0.0283	0.0266
k	κκ					0.0013	-	-	-	0.0000	-	-	-
ks	ξ	0.0086	0.0077	0.0086	0.0048	1.0000	1.0000	1.0000	1.0000	0.0086	0.0077	0.0086	0.0048

APPENDIX B: (Continued)

Ph	Gr	Phoneme probability				PGC probability				Sonograph probability			
		Corpus Size				Corpus Size				Corpus Size			
		4.10 K	1.66 K	0.70 K	0.21 K	4.10 K	1.66 K	0.70 K	0.21 K	4.10 K	1.66 K	0.70 K	0.21 K
l	λ	0.0355	0.0353	0.0349	0.0375	0.9325	0.9184	0.8889	0.8065	0.0331	0.0324	0.0310	0.0302
l	λλ					0.0675	0.0816	0.1111	0.1935	0.0024	0.0029	0.0039	0.0073
m	μ	0.0341	0.0331	0.0341	0.0399	0.9755	0.9783	0.9756	0.9697	0.0332	0.0324	0.0332	0.0387
m	μμ					0.0245	0.0217	0.0244	0.0303	0.0008	0.0007	0.0008	0.0012
n	v	0.0499	0.0478	0.0402	0.0423	0.9913	0.9849	0.9862	1.0000	0.0495	0.0471	0.0396	0.0423
n	vv					0.0087	0.0151	0.0138	-	0.0004	0.0007	0.0006	-
ɲ	vi	0.0022	0.0020	0.0019	0.0012	0.7333	0.6316	0.4286	-	0.0016	0.0012	0.0008	-
ɲ	l					0.1000	0.2632	0.2857	1.0000	0.0002	0.0005	0.0006	0.0012
ɲ	oi					0.0667	-	-	-	0.0001	-	-	-
ɲ	voi					0.0500	-	-	-	0.0001	-	-	-
ɲ	vvi					0.0333	0.0526	0.1429	-	0.0001	0.0001	0.0003	-
ɲ	u					0.0167	0.0526	0.1429	-	0.0000	0.0001	0.0003	-
ŋ	γ	0.0001	-	-	-	1.0000	-	-	-	0.0001	-	-	-
o	o	0.0554	0.0529	0.0562	0.0641	0.7407	0.7456	0.7143	0.7547	0.0410	0.0395	0.0402	0.0484
o	ω					0.2593	0.2544	0.2857	0.2453	0.0144	0.0135	0.0161	0.0157
'o	ó	0.0265	0.0280	0.0338	0.0459	0.7104	0.7500	0.7295	0.7105	0.0189	0.0210	0.0247	0.0326
'o	ώ					0.2814	0.2316	0.2459	0.2895	0.0075	0.0065	0.0083	0.0133
'o	ω					0.0068	0.0147	0.0164	-	0.0002	0.0004	0.0006	-
'o	o					0.0014	0.0037	0.0082	-	0.0000	0.0001	0.0003	-
p	π	0.0353	0.0392	0.0457	0.0508	0.9908	0.9869	0.9879	1.0000	0.0350	0.0387	0.0452	0.0508
p	ππ					0.0092	0.0131	0.0121	-	0.0003	0.0005	0.0006	-
ps	ψ	0.0033	0.0031	0.0025	0.0012	1.0000	1.0000	1.0000	1.0000	0.0033	0.0031	0.0025	0.0012
r	ρ	0.0580	0.0582	0.0615	0.0532	0.9950	0.9965	1.0000	1.0000	0.0577	0.0580	0.0615	0.0532
r	ρρ					0.0050	0.0035	-	-	0.0003	0.0002	-	-
s	σ	0.0775	0.0748	0.0729	0.0774	0.6183	0.5742	0.5323	0.5156	0.0479	0.0430	0.0388	0.0399
s	ς					0.3709	0.4107	0.4525	0.4844	0.0287	0.0307	0.0330	0.0375
s	σσ					0.0108	0.0151	0.0152	-	0.0008	0.0011	0.0011	-

APPENDIX B: (Continued)

Ph	Gr	Phoneme probability				PGC probability				Sonograph probability			
		Corpus Size				Corpus Size				Corpus Size			
		4.10 K	1.66 K	0.70 K	0.21 K	4.10 K	1.66 K	0.70 K	0.21 K	4.10 K	1.66 K	0.70 K	0.21 K
t	τ	0.0499	0.0486	0.0485	0.0665	1.0000	1.0000	1.0000	1.0000	0.0499	0.0486	0.0485	0.0665
ts	τσ	0.0018	0.0015	0.0019	0.0012	1.0000	1.0000	1.0000	1.0000	0.0018	0.0015	0.0019	0.0012
u	ου	0.0155	0.0145	0.0150	0.0109	1.0000	1.0000	1.0000	1.0000	0.0155	0.0145	0.0150	0.0109
'u	ού	0.0124	0.0115	0.0080	0.0048	0.9912	0.9911	0.9655	1.0000	0.0123	0.0114	0.0078	0.0048
'u	ου					0.0088	0.0089	0.0345	-	0.0001	0.0001	0.0003	-
v	β	0.0145	0.0145	0.0130	0.0036	0.7995	0.8652	0.8723	1.0000	0.0116	0.0125	0.0114	0.0036
v	υ					0.1128	0.0638	0.0638	-	0.0016	0.0009	0.0008	-
v	ύ					0.0877	0.0709	0.0638	-	0.0013	0.0010	0.0008	-
x	χ	0.0125	0.0130	0.0139	0.0036	1.0000	1.0000	1.0000	1.0000	0.0125	0.0130	0.0139	0.0036
λ	λι	0.0025	0.0026	0.0025	0.0012	0.8529	0.7600	0.8889	1.0000	0.0021	0.0020	0.0022	0.0012
λ	λει					0.1029	0.1600	0.1111	-	0.0003	0.0004	0.0003	-
λ	λλι					0.0441	0.0800	-	-	0.0001	0.0002	-	-
z	ζ	0.0108	0.0076	0.0050	0.0024	0.7710	0.8378	0.9444	1.0000	0.0083	0.0064	0.0047	0.0024
z	σ					0.2290	0.1622	0.0556	-	0.0025	0.0012	0.0003	-
ε	ε	0.0616	0.0601	0.0551	0.0532	0.9111	0.9060	0.8995	0.8636	0.0561	0.0545	0.0496	0.0459
ε	αι					0.0889	0.0940	0.1005	0.1364	0.0055	0.0057	0.0055	0.0073
'ε	έ	0.0277	0.0311	0.0321	0.0363	0.8480	0.8515	0.8621	0.8333	0.0235	0.0265	0.0277	0.0302
'ε	εί					0.0826	0.0924	0.0776	0.0333	0.0023	0.0029	0.0025	0.0012
'ε	ε					0.0682	0.0528	0.0517	0.1000	0.0019	0.0016	0.0017	0.0036
'ε	αι					0.0013	0.0033	0.0086	0.0333	0.0000	0.0001	0.0003	0.0012
θ	θ	0.0108	0.0097	0.0072	0.0097	1.0000	1.0000	1.0000	1.0000	0.0108	0.0097	0.0072	0.0097

REFERENCES

- Berndt, R., Reggia, J., & Mitchum, 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.
- Chliounaki, K., & Bryant, P. (2002). Construction and Learning to spell. *Cognitive Development*, 17, 1489-1499.
- Coltheart, M., (2006). Dual route and connectionist models of reading: An overview. *London Review of Education*, 4, 5-17.
- 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.
- Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., Van Daal, V. H., Polyzoe, N., et al. (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.
- Gontijo, P. F., Gontijo, I., & Shillcock, R. (2003). Grapheme-phoneme probabilities in British English. *Behavior Research Methods, Instruments & Computers*, 35 , 136-157.
- 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.
- Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., et al. (2000, 31 May-2 June). Design and implementation of the online ILSP corpus. In *Proceedings of the second international conference of language resources and evaluation (LREC)* (Vol. 3, pp. 1737-1740). Athens, Greece.
- Hatzigeorgiu, N., Mikros, G., & Carayannis, G. (2001). Word Length, Word Frequencies and Zipf's Law in the Greek Language. *Journal of Quantitative Linguistics*, 8 , 175 – 185.
- Karantzola, E., Kurdi, K., Spanelli, T., & Tsiagkani, Th. (2008). *Glossa A'Dimotikou, Volumes 1 & 2*. Athens: OEDB.
- Katz, L., & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In R. Frost, & L. Katz, *Orthography, phonology, morphology, and meaning* (pp. 67-84). Amsterdam: Elsevier Science Publishers.
- Ktori, M., van Heuven, W. J., & Pitchford, N. J. (2008). GreekLex: A lexical database of Modern Greek. *Behavior Research Methods*, 40 , 773-783.

- Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition*, 63 , 315–334.
- Loizidou-Ieridou, N. (2007). *Literacy development and reading difficulties in Greek-speaking Cypriot children aged between 6 and 11 years*. Unpublished PhD thesis, University of Essex.
- Loizidou-Ieridou, N., Masterson, J., & Hanley, R. (2009). *Spelling development in 6-11 year old Greek-speaking Cypriot children*. Manuscript submitted for publication.
- Mackridge, P. (1985). *The Modern Greek Language: A descriptive analysis of standard Modern Greek*. Oxford: Oxford University Press.
- Martindale, C., Gusein-Zade, S. M., McKenzie, D., & Borodovsky, M. Y. (1996). Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics*, 3 , 106-112.
- 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.
- Mikros, G., Hatzigeorgiu, N., & Carayannis, G. (2005). Basic quantitative characteristics of the Modern Greek language using the Hellenic National Corpus. *Journal of Quantitative Linguistics*, 12 , 167-184.
- Oney, B., & Goldman, S. (1984). Decoding and comprehension skills in Turkish and English: Effects of regularity of grapheme-phoneme correspondences. *Journal of Educational Psychology*, 76 , 557-568.

- Palmer, M. (2002). *A Quick Overview of the History of the Greek Language*.
Retrieved March 18, 2009, from <http://www.greek-language.com>
- 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(1): 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.
- 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.
- Porpodas, C. D. (1999). Patterns of phonological and memory processing in beginning readers and spellers of Greek. *Journal of Learning Disabilities*, 32 , 406-416.
- Porpodas, C. D. (2006). Literacy acquisition in Greek: Research review of the role of phonological and cognitive factors. In *Handbook of orthography and literacy* (pp. 189-199). Mahwah, NJ: Erlbaum.
- Protopapas, A., & Vlahou, E. (in press). A comparative quantitative analysis of Greek orthographic transparency. *Behavior Research Methods*.
- Pring, J. T. (1986). *The Oxford dictionary of modern Greek*. Oxford: Oxford University Press.
- 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. (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. (2009). Feedforward, -backward, and neutral transparency measures for British English. *Behavior Research Methods*, *41*, 220-227; doi:10.3758/BRM.41.1.220.
- 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.
- 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.
- Zachos, J. (1991). *Language and linguistic material*. Athens: CPR.
- 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.

Zipf, G. K. (1936). *The Psycho-Biology of Language: An Introduction to Dynamic Psychology*. London: Routledge.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley Press.

Table 1*Metric calculations for sonograph /i/-<ει>*

	Frequency	Total	Probability
Sonograph	437	27585	0.0158
Grapheme	447	27585	0.0162
Phoneme	2412	27585	0.0874
Phoneme-grapheme PGC	437	2412	0.1812
Grapheme-phoneme GPC	437	447	0.9776

Table 2*Correlations among five word metrics for Corpus sizes 0.21K – 4.10K*

	1	2	3	4	5	6	7	8
1. Corpus 4.10K: G	-	1.00 **	.98 **	.96 **	-.22 *	-.24 *	-.26 *	-.25 *
2. Corpus 1.66K: G		-	.99 **	.97 **	-.23 *	-.24 *	-.26 *	-.25 *
3. Corpus 0.70K: G			-	.97 **	-.19	-.21	-.24 *	-.23
4. Corpus 0.21K: G				-	-.15	-.16	-.17	-.18
5. Corpus 4.10K: GP					-	.99 **	.96 **	.94 **
6. Corpus 1.66K: GP						-	.98 **	.95 **
7. Corpus 0.70K: GP							-	.96 **
8. Corpus 0.21K: GP								-
9. Corpus 4.10K: P								
10. Corpus 1.66K: P								
11. Corpus 0.70K: P								
12. Corpus 0.21K: P								
13. Corpus 4.10K: SG								
14. Corpus 1.66K: SG								
15. Corpus 0.70K: SG								
16. Corpus 0.21K: SG								
17. Corpus 4.10K: PG								
18. Corpus 1.66K: PG								
19. Corpus 0.70K: PG								
20. Corpus 0.21K: PG								

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.

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

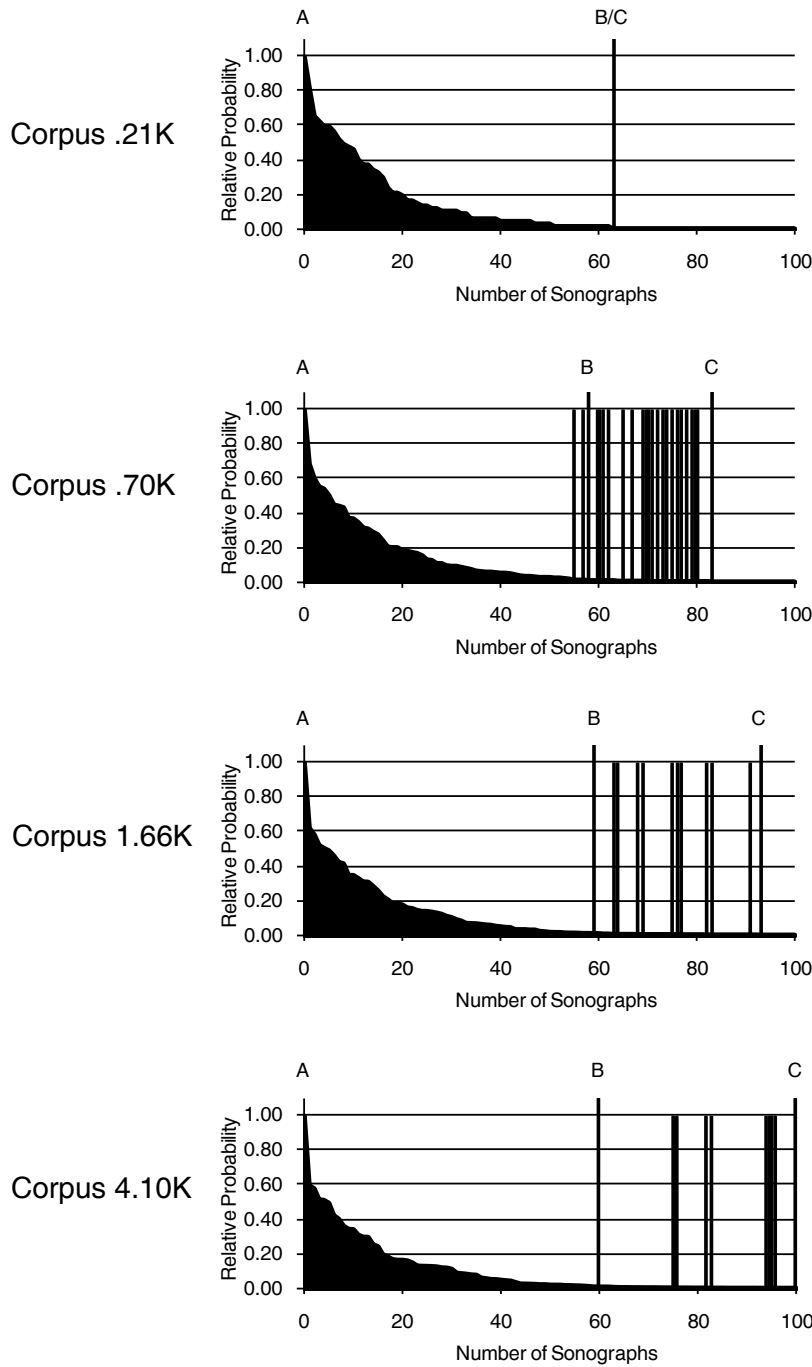


Figure 1. Sonograph probability profiles for four corpus sizes. For comparative purposes relative probability values are expressed as a proportion of the highest frequency sonograph. Position B indicates relative probabilities $< .01$. Position C = total number of sonographs for the corpus. Bars in area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

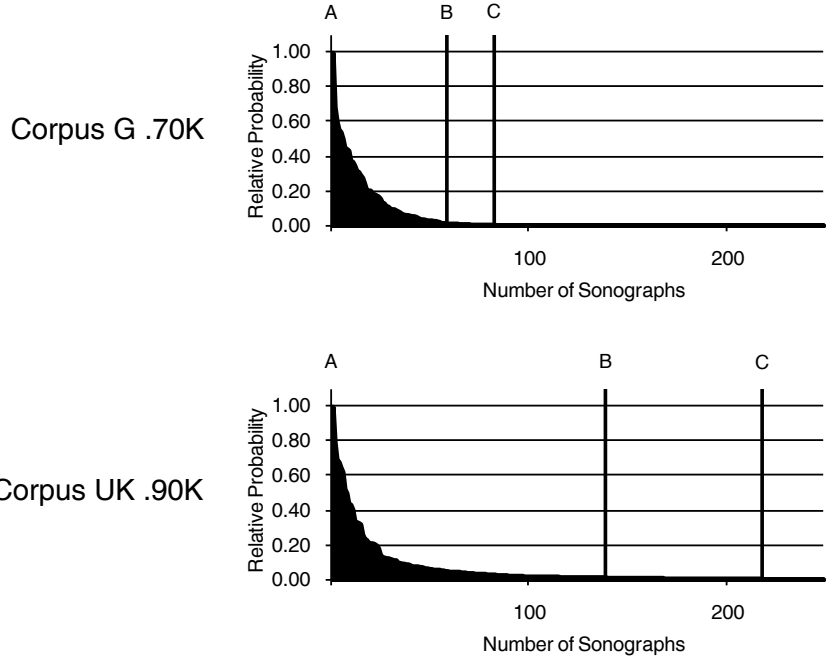
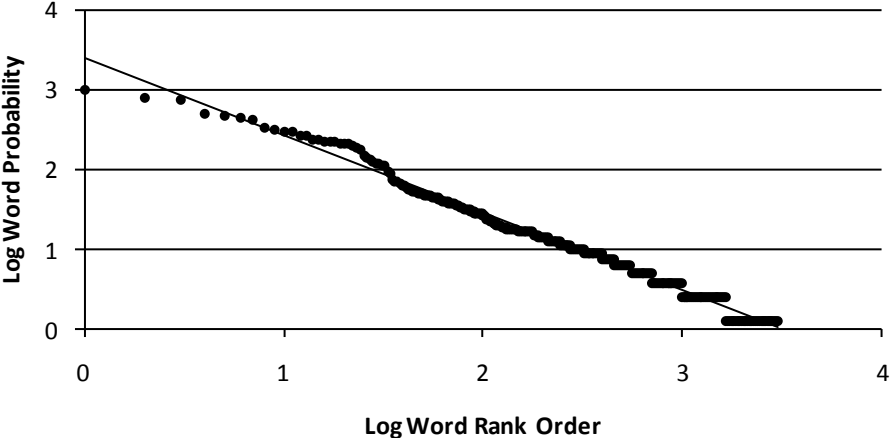
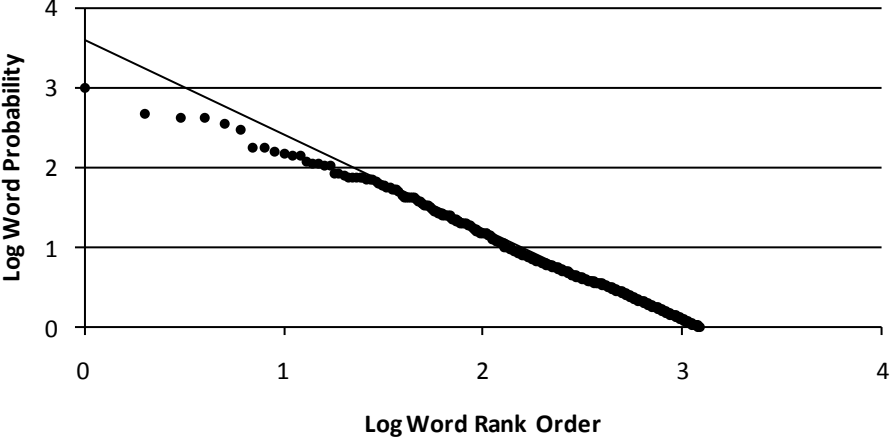


Figure 2. Sonograph probability profiles for G (Greek) and UK (British English, based on Spencer, 2009) for corpora of similar size. For comparative purposes relative probability values are expressed as a proportion of the highest frequency sonograph. Position B indicates relative probabilities < .01. Position C = total number of sonographs for the corpus.

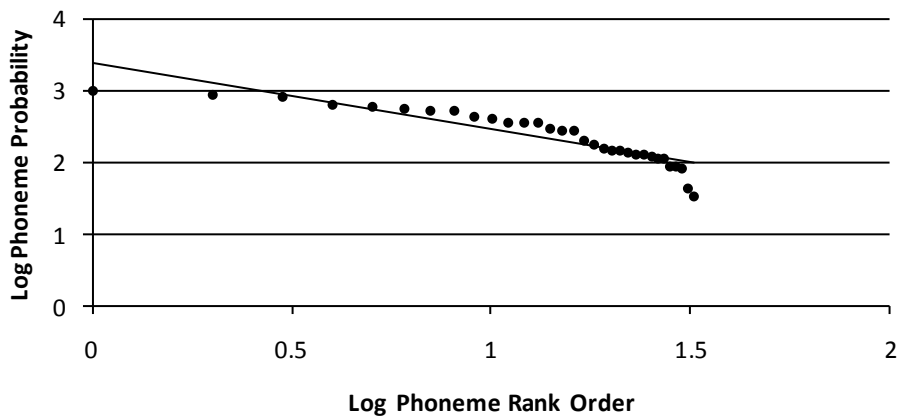
Slope (Greek) = -0.97; R² = 0.98



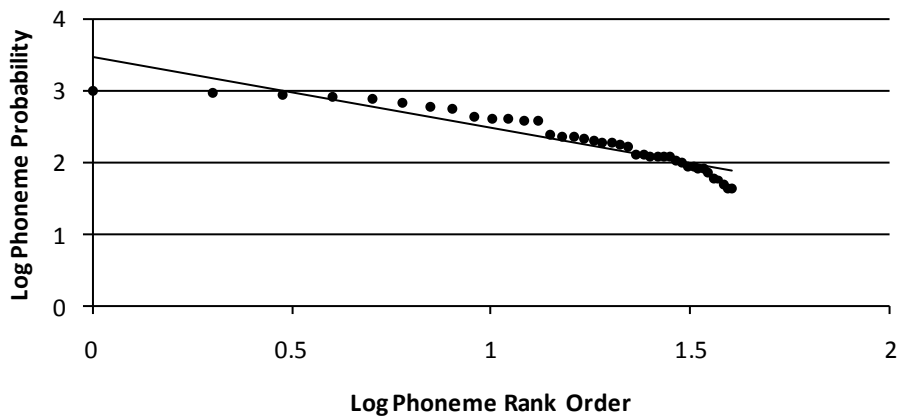
Slope (English) = -1.19; R² = 0.98



Slope (Greek) = -0.92; R² = 0.82



Slope (English) = -0.99; R² = 0.88



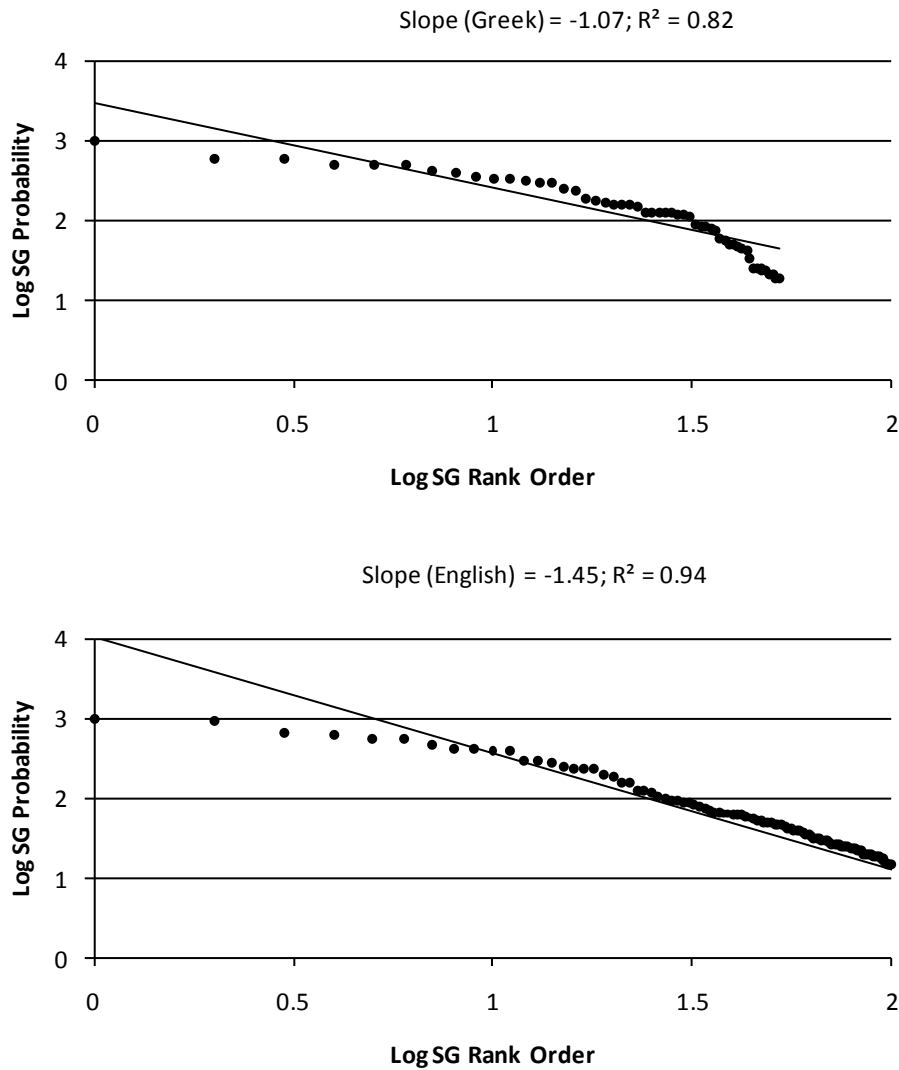


Figure 3. Log-log plots for Greek and English words, phonemes, and sonographs. For comparative purposes frequency values are expressed as a proportion of 1,000 occurrences. Word frequencies are based on ranks 1 to 3,000; phoneme and sonograph frequencies are based on 99% cumulative values to prevent excessive tail end distortion.

¹ For comparative purposes probabilities are expressed relative to the most frequent sonograph.