A comparative quantitative analysis of Greek orthographic transparency

Athanassios Protopapas
Institute for Language & Speech Processing

Eleni L. Vlahou
Institute for Language & Speech Processing
and University of Crete

Orthographic transparency refers to the systematicity in the mapping between orthographic letter sequences and phonological phoneme sequences, in both directions: for reading and spelling. Measures of transparency previously used in the analysis of orthographies of other languages include regularity, consistency, and entropy. However, previous reports are typically hampered by severe restrictions such as using only monosyllables or only word-initial phonemes. Greek is sufficiently transparent to allow complete sequential alignment between graphemes and phonemes, therefore permitting full analyses at both letter and grapheme level using every word in its entirety. Here we report multiple alternative measures of transparency using both type and token counts and compare to estimates for other languages. We discuss the problems stemming from restricted analysis sets and the implications for psycholinguistic experimentation and computational modeling of reading and spelling.

Alphabetic orthographies differ in their degree of transparency, that is, in the systematicity of the mapping between letter sequences and phoneme sequences. Inconsistencies in the sound-spelling mappings arise when single orthographic units have multiple pronunciations or single phonological units have multiple spellings. Quantitative assessments of such ambiguities have been carried out in several languages with alphabetic orthographic systems, both in the "feedforward" direction, that is, from orthography to phonology, as needed for reading aloud printed words, and in the "feedback" direction, from phonology to orthography, as needed for spelling (e.g., Borgwaldt, Hellwig, & De Groot, 2005, 2004; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995; Ziegler, Stone, & Jacobs, 1997; Ziegler, Jacobs, & Stone, 1996).

The study of orthographic transparency is important for theoretical and practical reasons, as ambiguous mappings have been found to affect reading and spelling performance (Spencer, 2007, submitted). For example, feedforward-inconsistent orthographic units (i.e. letter sequences that can be pronounced in more than one way) slow down word naming (Burani, Barca, & Ellis, 2006; Jared, 2002; Treiman et al., 1995), while ambiguities in the feedback direction affect spelling performance (Lété, Peereman, & Fayol, 2008; Burt & Blackwell, 2008). A counterintuitive finding is that feedback inconsistency also affects reading. That is, words with predictable pronunciation but unpredictable spelling are read and recognized more slowly than words with predictable spelling (Grainger & Ziegler, 2008; McKague, Davis, Pratt, & Johnston, 2008). However, in a review of studies on feedback inconsistency effects, Kessler, Treiman, and Mullennix (2008) pointed out a number of methodological shortcomings that need to be addressed before a final conclusion can be reached. To pursue these issues in additional languages, detailed quantification of orthographic consistency in both directions is necessary.

Ambiguity does not affect all sublexical units equally. In more opaque orthographies, smaller units tend to be less consistent than larger units (Ziegler & Goswami, 2005). For example, graphemes[1] are less consistent than orthographic bodies (spellings of a syllabic rime, i.e., of the nuclear vowel and any consonants that follow it) in English monosyllables (Treiman et al., 1995). There is thus a functional pressure for readers to develop both small-unit and large-unit recoding strategies. As the grain size grows, the number of distinct orthographic units rises. This "granularity" problem is more pervasive in opaque orthographies, whereas readers of more transparent orthographies can focus on finer grain sizes (Ziegler & Goswami, 2005). This assertion remains to be substantiated with specific estimates of the transparency and granularity of specific orthographic systems.

In the present work we address two issues: First, we provide a systematic quantitative exposition of transparency in the Greek orthography. The goals of this presentation are

We thank Aimilios Chalamandaris and the ILSP text-to-speech group for providing the text corpora and reference phonetic transcriptions, Elina Nomikou and Stella Drakopoulou for checking CiV words, Aikaterini Pantoula for help with the monosyllables, and Efthymia C. Kapnoula for help with the GPC rules.

The quantification of Greek orthographic transparency is part of a larger ongoing effort at ILSP to provide psycholinguistic resources for the Greek language. Data tables, code implementing graphophonemic conversion by rule, and other material may be downloaded from `speech.ilsp.gr/iplr/`.

Correspondence regarding this article may be sent to A. Protopapas at ILSP, Artemidos 6 & Epidavrou, GR-151 25 Maroussi, Greece; e-mail: protopap@ilsp.gr.

[1] A grapheme is the written representation of one phoneme, that is, a letter or group of letters that correspond to a single phoneme (Coltheart, 1978; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001).

to support researchers working on Greek with information relevant to stimulus selection and experimental design, and to provide researchers working in other orthographies with a comparison reference. Second, we use Greek as a test case to examine and evaluate a multitude of approaches to the quantification of orthographic transparency. By comparing and contrasting various alternative methods that have been proposed in the literature we are able to test certain assumptions underlying them and indicate weaknesses that may limit their applicability. Thus, our analysis suggests, directly or indirectly, ways towards improving the quantification of orthographic transparency for future research.

In the following sections, we first present existing approaches to the quantification of orthographic transparency in different alphabetic writing systems, discussing their methodological strengths and weaknesses. We then introduce the most important aspects of the Greek orthography. We report analyses of the transparency of Greek orthography, comparing and contrasting calculations of regularity and consistency, based on a word-form list from a representative corpus of contemporary written Greek texts. We identify the grapheme-phoneme level of analysis as the most appropriate for Greek and we present statistics relating individual phonemes and graphemes, as well as an ordered set of rules maximally capturing graphophonemic transcription. Finally, we discuss the implications of our findings for cross-linguistic evaluation of orthographic transparency, for learning to read and write in Greek, as well as for modeling reading Greek.

Quantitative indices of orthographic transparency

Regularity

The *regularity* approach assumes the theoretical position that there are "regular" mappings, governed by symbolic transcription rules, and "irregular" mappings, which violate the rules. Under this framework, the problem consists in the specification of a set of rules that relates individual graphemes to the corresponding phonemes (for the feedforward direction; or the reverse for the feedback direction). In cases where the mapping deviates from one-to-one, for example when a single grapheme can have multiple pronunciations, the most frequent mapping is considered regular and the others irregular. Regular words are words whose pronunciation or spelling is correctly produced by the grapheme-phoneme correspondence rules of a language, while irregular or exception words are words whose pronunciation or spelling cannot be predicted from these rules (Coltheart et al., 2001; Ziegler, Perry, & Coltheart, 2003). Regularity is thus conceptualized as a categorical distinction (Zevin & Seidenberg, 2006).

Ziegler, Perry, and Coltheart (2000) compared the degree of regularity of English and German by examining the pronunciations produced by the sublexical reading route of the "dual route cascaded" model (DRC; Coltheart et al., 2001) when presented with monosyllabic words from each language. By definition, the pronunciation produced by the sublexical route of the DRC will be wrong for every irregular word as it will tend to regularize it, according to the specified conversion rules of the language. Therefore, an evaluation of the model's performance provides an index of the degree of regularity of the language. Using this rule-based approach, Ziegler et al. found that the percentage of correct rule application is 90.4% for German as compared to 79.3% for English monosyllabic words. Because this approach has not yet been applied to polysyllabic words, it is unknown whether these are valid estimates for more representative samples.

Consistency

As an alternative to regularity, the *consistency* approach forgoes the notion of rules. Consistency refers to the (lack of) variability in the correspondences between the phonological and orthographic units of a language. For example, the consistency of a grapheme (in the feedforward direction; or a phoneme, in the feedback direction) decreases as the number and relative frequency of its corresponding alternative pronunciations (or spellings, respectively) increases (e.g., Lété et al., 2008; Perry, Ziegler, & Coltheart, 2002). Consistency computations can be performed at the grapheme-phoneme level or at larger grain sizes and can be dichotomous or graded. In dichotomous analyses, a word (or smaller-size unit) is considered consistent when there is only one possible mapping for it, or inconsistent when alternative mappings are possible. In graded analyses, the measure of consistency quantifies ambiguity by taking into account the relative frequency of alternative mappings and is expressed as the proportion of dominant mappings over the total number of occurrences of the base unit of analysis.

Consistency estimation based on grain sizes larger than the phoneme-grapheme is a common approach for the English language because taking into account larger parts of the syllable reduces ambiguity (Kessler & Treiman, 2001; Peereman & Content, n.d.; Treiman et al., 1995; Ziegler et al., 1997). Treiman et al. (1995) performed statistical analyses of the spelling-to-sound relations of English monosyllabic words with CVC (consonant, vowel, consonant) structure. They found that the vowel unit is highly inconsistent when examined individually or in combination with the initial consonant (head). But when the vowels were examined together with the following consonants (forming rimes), the English orthography appeared to be much more consistent. Kessler and Treiman (2001) extended the analysis in both directions (reading and spelling). They introduced conditional consistencies and permutation tests of significance to address questions such as whether the rime is processed as a whole or whether the influence of the intrarime units is symmetrical (i.e., whether the vowel and the coda improve the consistency of each other). They concluded that rimes are not processed as individual units. Rather, the basic processing seems to occur at a phonemic-graphemic level that takes into account the context in which each phoneme-grapheme is found.

Based on the rime-body level and using a dichotomous classification, Ziegler et al. (1996, 1997) performed bidi-

rectional statistical analyses of French and English monosyllabic monomorphemic words. A word was considered consistent if there was a one-to-one correspondence between the word's spelling body and its phonological body. The results showed that the degree of inconsistency differs between these languages and within each language between the feedforward and feedback direction. From the spelling point of view, both English and French can be described as opaque languages, as 79.1% of the French words and 72.3% of the English words were feedback inconsistent. From the reading perspective, 12.4% of the French words and 30.7% of the English words were feedforward inconsistent. Therefore both languages are more consistent in the feedforward direction but the asymmetry is much higher for French.

A severe limitation of the aforementioned studies is the computation of consistency values based solely on monosyllabic words. Consistency measures based on a nonrandom subset of words in a language may not constitute reliable estimates of the sound-spelling relations of this language as a whole, insofar as polysyllables may have a different structure from monosyllables of may be otherwise biased (Borgwaldt et al., 2004; Kessler & Treiman, 2001; Lété et al., 2008). Moreover, restriction of computations to monosyllables limits the potential for cross-linguistic comparisons, because languages differ in the proportion of words that are monosyllables and possibly in the representativeness of monosyllables with respect to the full spectrum of orthographic mappings. Finally, analyses at the rime-body level may only be justified for languages such as English, where smaller grain sizes would lead to very low consistency estimates. In more transparent orthographies, grapheme-phoneme mappings are more consistent across the entire vocabulary, highlighting the importance of selecting a cross-linguistically appropriate level of analysis. These issues are addressed in the present study by a comparative analysis at different grain sizes, using an unabridged lexical database.

Entropy

The calculation of consistency as the proportion of majority mappings suffers from the inability to discriminate between cases with many and few alternatives, and between nondominant mappings with substantial and negligible proportions. For example, in Greek, the phoneme [g][2] can be spelled as either ⟨γκ⟩ (85.5%) or ⟨γγ⟩ (14.5%). The phoneme [ç] can be spelled as ⟨χ⟩ (85.0%), ⟨οι⟩ (7.0%), ⟨ι⟩ (6.9%), or as ⟨ει⟩, ⟨χι⟩, ⟨χει⟩ and other combinations with very low probability (less than 1% each). According to the consistency index, these phonemes are equally consistent at about 85%. However, their mappings are not equally unpredictable, because there is only one significant nondominant option for [g] but two for [ç] (among several minor ones). The consistency index cannot express this difference in the ambiguity of mapping between the two phonemes.

This shortcoming can be remedied by resorting to entropy, a more sophisticated index of consistency, which assesses the same underlying concept but can take into account the complications arising from the entire distribution of mappings and not only the dominant ones. Entropy is an information-theoretic notion that quantifies uncertainty (i.e., lack of information; Shannon, 1948, 1950). In the context of graphophonemic transparency, entropy quantifies ambiguity in the prediction of letters or graphemes by phonemes and vice versa. If a given grapheme (phoneme) maps unambiguously to a specific phoneme (grapheme), then the mapping is absolutely certain, that is, there is no ambiguity, hence the corresponding entropy is zero. Entropy is high for graphemes (phonemes) with many alternative pronunciations (spellings), especially when there is not a single dominant mapping.

For any unit of orthographic (or phonological) representation that maps onto n phonological (orthographic) alternatives with probability p_i for the ith alternative, its entropy (H) is calculated as the negative sum, over the alternative mappings, of the products of each probability times its logarithm:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Using base-2 logarithms for the calculation results in a quantification expressed in bits. For example, the entropy of the [g] phoneme mentioned above would be $-[0.855 \times \log_2(0.855) + 0.145 \times \log_2(0.145)] = 0.597$ bits. To calculate the total entropy associated with an entire set of units (letters, graphemes, or phonemes), the contribution of each individual unit is weighed by its relative frequency of occurrence before it is added to the total. Calculated in the same way, the entropy of [ç] equals 0.827 bits.

Borgwaldt et al. (2004, 2005) performed analyses of entropy for English, Dutch, German, French, Hungarian, Italian and Portuguese. To overcome the limitations of restricting analysis to monosyllables, they focused on word-initial sound-spelling correspondences using all the words in each language. The restriction to word-initial mappings was dictated by severe constraints on defining and segmenting the units of analysis in a practical and cross-linguistically uniform manner, due to difficulties in identifying the borders of graphemes within words, at least in some languages. Borgwaldt et al. showed that none of the orthographies examined approached the ideal one-to-one mapping between letters and sounds and that word-initial letter entropy is significantly correlated with word naming latency in Italian, Dutch, and English, three languages varying widely in orthographic transparency. However, it remains to be empirically substantiated whether word-initial mappings constitute an unbiased representative sample of all mappings. This is an important issue we address empirically in the present study.

[2] Greek letters and graphemes in text are shown within angle brackets, to avoid confusion. Phonetic symbols refer to broadly transcribed phonetic realizations, not to any presumed underlying phonological representations, and are shown within square brackets.

Type vs. token counts

An issue warranting further scrutiny concerns the nature of the counts entering the calculations of transparency indices. In principle, consistency counts of sublexical units can be performed based on either word-form or lemma databases and on either type or token frequency counts. The difference in database contents concerns items related via inflectional morphology. That is, word-form databases contain morphological variants of the same lemma whereas lemma databases contain only a "base" form. Type counts of sublexical units consist in the number of items (word forms or lemmas) that contain the unit in question. Token counts are calculated by summing the frequency (number of occurrences in a text corpus) of the items that contain the unit.

The optimal choice between lemmata and word forms depends on the researcher's theoretical assumptions and research goals (Hofmann, Stenneken, Conrad, & Jacobs, 2007). However, in lemma databases, the sublexical units of the base form that are not present in the inflected forms are overestimated, whereas units associated with inflections are underestimated. This led Hofmann et al. to recommend using word-form databases, especially when assessing language in its natural, inflected form.

The choice between type vs. token measures remains controversial, as both seem to be independently associated with lexical processing. Conrad, Carreiras, and Jacobs (2008) showed that a token-based measure of syllable frequency was associated with an inhibitory effect on lexical access in a lexical-decision task, whereas a type-based measure was associated with a facilitative effect. To account for the contradictory findings, Conrad et al. proposed that the two kinds of measures are related to different processing stages during visual word recognition. In contrast, Moscoso del Prado Martín, Ernestus, and Baayen (2004) have proposed that a common, token-based mechanism can account for both token- and type-based effects. They modeled Dutch past tense formation with a simple recurrent network which exhibited token-based frequency effects and type-based analogical effects that closely matched the behavior of human participants.

In the aforementioned studies employing entropy, Borgwaldt et al. (2004) used word form types whereas Borgwaldt et al. (2005) used lemma types. Neither option corresponds to the cumulative experience of readers and spellers with the graphophonemic mappings because the frequency of occurrence of each word in written and spoken language was not taken into account. Insofar as the frequency of occurrence of a word is among the strongest predictors of how quickly it can be recognized or read aloud (Balota, Yap, & Cortese, 2006), it might be preferable to measure indices of transparency in terms of tokens instead of type counts, using word forms weighted by the number of their occurrences in a representative text or speech corpus. A more conservative approach would take into account both type and token frequency counts, following the recommendation of Hofmann et al. (2007).

The Greek orthography

There are 32 phonemes in modern Greek at the level of broad phonetic transcription (discounting idiolectal and optional variation), of which 5 are vowels. These are written with 24 letters (plus one final-only form), of which 7 correspond to vowels in isolation. Most words can be read correctly on the basis of the letter sequence alone, without the need for morphological or lexical information. However, spelling is more complicated, because it is determined not only by phonological identity but also by morphological type, for grammatical inflections, and by the historical origin, for word stems tracing back to ancient Greek. There are two letters for the vowel [o] (⟨ο⟩, ⟨ω⟩) and two ways of spelling [ɛ] (⟨ε⟩ and the digraph ⟨αι⟩); [u] is spelled with the digraph ⟨ου⟩. There are six ways to spell [i] (⟨ι⟩, ⟨η⟩, ⟨υ⟩, ⟨ει⟩, ⟨οι⟩, ⟨υι⟩). As there are fewer consonant letters than phonemes, several consonants are spelled with digraphs. For example, all voiced stops are spelled with a combination of the letters for the corresponding unvoiced stop and the nasal at the same place of articulation: [m]→⟨μ⟩, [p]→⟨π⟩, [b]→⟨μπ⟩. Palatal consonants are spelled with the letter for the corresponding velar consonant and one of the [i] graphemes. Despite these and other complications, learning to read Greek is considered to be relatively easy, and online resources are available providing the necessary information (e.g., "The Greek Alphabet" at `http://www.xanthi.ilsp.gr/filog/ch1/alphabet/alphabet.asp`).

The Greek orthography is commonly characterized as "transparent" or "shallow" despite a dearth of relevant quantitative data. In the classification of several European orthographic systems by Seymour, Aro, and Erskine (2003), in the context of a cross-linguistic study on early stages of learning to read, Greek occupied the second position, in order of decreasing orthogaphic transparency for reading, in the group of languages with relatively simple syllabic structures, after Finnish. However, There seems to be a great asymmetry in the transparency of Greek orthography between the feedforward (reading) and feedback (spelling) direction. According to Porpodas (2006), in the feedforward direction "[t]he Greek spelling system... [approaches] a 1:1 relationship between graphemes and phonemes... and can be characterized as a *shallow* orthography in which, as a rule, pronunciation is predictable from print", whereas in the feedback direction "Greek is *phonologically opaque* as there is a one-to-many phoneme-grapheme mapping and therefore spelling cannot always be predictable from phonology" (p. 192, emphasis in original).

Consistent with the notion of rule-based predictability, Petrounias (2002) has listed a set of rules for each direction of conversion. However, according to Petrounias, some of the rules apply only in words of the vernacular and are often violated in words of literary or learned origin. As the origins of each word are not necessarily clear to contemporary Greek speakers, this diachronically systematic distinction constitutes a source of synchronic inconsistency. Petrounias lists several cases of mappings between phonemes and letters, which can be generally classified into one-to-one, one-to-

many, and many-to-one. Deviations from one-to-one include both digraphs/diphones (i.e., phonemes spelled with two or more letters, such as [b]→⟨μπ⟩, or single letters pronounced as two phonemes, such as ⟨ξ⟩→[ks]) and context-dependent transcriptions (i.e., phonemes spelled differently depending on adjacent phonemes, e.g., [z]→⟨σ⟩ before [v] vs. [z]→⟨ζ⟩ before [o], or letters pronounced differently depending on adjacent letters, e.g., ⟨κ⟩→[k] before ⟨ο⟩ vs. ⟨κ⟩→[c] before ⟨ε⟩).

Once the assumption of rules is made, the notion of regularity becomes relevant. Exceptions are only possible when rules are defined to which they do not conform. In Greek there are some clear exceptions, typically recent loans, which violate the rules and are pronounced similar to their foreign origin (at least by educated speakers). For example, Greek has no way to spell [mp] or [nt] (and these clusters do not occur within native words) because, as noted above, the corresponding letter combinations (⟨μ⟩ for [m] and ⟨π⟩ for [p]; ⟨ν⟩ for [n] and ⟨τ⟩ for [t]) are used as digraphs for the voiced stops. Therefore, words such as σαμπάνια ("champagne") and κάμπινγκ ("camping") that have entered the vocabulary recently, are properly pronounced [samˈpaɲa] and [ˈkampiŋg], respectively, in violation of the grapheme-to-phoneme rules.

The most pervasive issue of inconsistency and irregularity in the feedforward direction concerns the general phenomenon of CiV, that is, the occurrence of an [i] grapheme preceded by a consonant and followed by a vowel. In every such case there are two possible pronunciations: one that includes the [i] and one that includes a palatal consonant and no [i]. In the former case the CiV is parsed into 3 graphemes (C, i, V) and the i grapheme is indeed pronounced [i], as in ήλιο ("helium") pronounced [ˈilio] and άδεια ("permission") pronounced [ˈaðia]. In the latter case the CiV is parsed either into 2 graphemes (Ci, V), the Ci part corresponding to the palatal consonant, as in ήλιο ("sun") pronounced [ˈiʎo], or into 3 graphemes (C, i, V), in which case the palatal consonant actually corresponds to the i grapheme, as in άδεια ("empty") pronounced [ˈaðja]. Note that, although homographs are used in this example to make the point most clearly, there are in fact very few homographs of this sort. For the vast majority of letter strings there is only one correct (i.e., word-forming) parsing of the CiV. Compare, for example, ντόπιο→[ˈdopço] vs. όπιο→[ˈopio] and ζήλεια→[ˈziʎa] vs. τέλεια→[ˈtɛlia].

Greek orthography marks lexical stress with a special diacritic. As noted by Petrounias (2002), the current spelling convention for Greek regarding the stress diacritic contains an element of inconsistency because its application depends on the number of syllables and not only on the presence of phonological stress. Specifically, monosyllables do not bear a diacritic, whereas every word with two or more syllables must bear a diacritic. This is phonologically appropriate in the majority of cases, because most monosyllables are grammatical words that attach themselves metrically to adjacent content words. Conversely, most polysyllables are content words and bear phonological stress, which is always correctly marked with the diacritic. However, there are exceptions: Monosyllabic content words, which bear phonological stress, are not marked with the diacritic due to the spelling convention, whereas disyllabic function words, which do not bear phonological stress, are written with a diacritic nonetheless (see Petrounias, 2002, pp. 533–534).

In addition to its metrical significance, the stress diacritic sometimes helps disambiguate graphophonemic mappings, because of spelling conventions concerning vowel digraphs. For example, ⟨ει⟩ and ⟨εί⟩ constitute graphemes and are pronounced [i], whereas ⟨έι⟩ includes two graphemes and is pronounced [ei]. A diaeresis diacritic is also available, to disambiguate single vowel graphemes from digraphs, so ⟨εϊ⟩ and ⟨εΐ⟩ are also bigraphemic, pronounced [ei].

In conclusion, there is a certain degree of complexity and some inconsistency in Greek spelling at the level of individual letters and phonemes, in both directions. In the present study we quantify it, comparing and contrasting results from calculations of regularity, consistency, and entropy using type and token counts from a word-form list derived from a representative corpus of contemporary written Greek texts. Because of the characteristics of Greek spelling, it is possible to calculate these indices for all words, regardless of their length, and for all letters and phonemes of every word. In this way we can address shortcomings of previous studies: Specifically, by applying all three approaches we can examine whether different metrics of transparency may lead to similar conclusions and predictions. Most importantly, we can critically assess the validity of assumptions that have led researchers to perform their analyses on restricted sets of words (e.g., monosyllables) or parts thereof (e.g., word beginnings or rimes).

Method

Text corpus

All analyses and counts were performed on a word list derived in 2006 from the Hellenic National Corpus (HNC; Hatzigeorgiu et al., 2000; `http://hnc.ilsp.gr`). This is an evolving corpus of a great variety of post-1990 widely circulated printed Greek texts including literary, journalistic, legal, and other texts from online news sources, newspapers, books, magazines, reports, proceedings, and brochures. The raw texts available at the time were tokenized into 31,363,642 white space-separated tokens and condensed into a list of 374,075 unique types with associated occurrence counts (frequency). Items (.4% of tokens) including any latin characters, numerals, or symbols were rejected, as were items (5.3% of tokens) not found in an electronic dictionary with 1,622,668 entries covering all possible morphological variants of inflected words ("Symfonia;" Stathis & Carayannis, 1999; `http://www.ilsp.gr/correct_eng.html`), resulting in a list of 217,664 unique word forms (types) accounting for total 29,557,090 occurrences (tokens). This word list was relatively free from spelling errors and contained few idiosyncratic items such as proper names, foreign words not quite integrated as loans in the Greek language, or very low frequency words unlikely to be found in the dictionary.

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 ς, a total of 36 letters were used in the analyses.

Phonetic transcription and post-processing

The list of orthographic types from the HNC was processed by a module producing phonetic representations of words that was developed for a Text-to-Speech project (Chalamandaris, Raptis, & Tsiakoulis, 2005). This module has been extensively validated, is known to produce highly accurate results, and has been in commercial and research use for several years. The resulting list of phonological types, corresponding to the HNC orthographic types, was post-processed to maximize uniformity by simplifying optional pronunciations that might result in unnecessary, and potentially misleading, complexity. Specifically, (a) all homorganic nasal obstruents preceding voiced stops were removed (e.g., [mb]→[b], [nd]→[d] etc.); (b) all [mpt] sequences were reduced to [mt]; and (c) all instances of [ɱ] were converted to [m]. These cases concern optional alternative pronunciations (phonologically and lexically nondistinctive) with variants freely alternating not only between dialects but also within dialects and talkers as a matter of sociolinguistic context or careful vs. relaxed articulation. In no instance are the simplified versions used in the subsequent analyses inappropriate, unusual, or otherwise marked. 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.

In addition, all types containing CiV sequences were identified and submitted to manual verification. The CiV pattern was found in 17.9% of the corpus types, amounting to 6.6% of the tokens. A custom software presented each of the 38,926 orthographic CiV-types individually while simultaneously playing out a synthesized pronunciation of the phonetic string derived from the grapheme-to-phoneme module. A listener indicated manually any errors in the transcription. Types for which both alternative pronunciations were acceptable (e.g., ⟨διάλογος⟩→[ðiˈaloɣos] or [ˈðjaloɣos]) were not modified. As a result of this procedure, a revised list of phonetic types was generated.

Grapheme alignment

Minimal experimentation revealed that it was always possible to align the set of phonemes making up each phonetic type with a set of graphemes making up the corresponding orthographic type, such that all phonemes and all letters were appropriately matched and none were left unassigned (no null phonemes or null graphemes). In other words, a strictly sequential grapheme-to-phoneme alignment is possible for Greek, fully accounting for all letters and phonemes, with the limitation that, because processing applies to individual word units, sandhi is effectively ignored. For example, here is the beginning of "the northern wind and the sun" fable, aligned at the grapheme-phoneme level:

ο	β	ο	ρ	ι	ά	ς	κι	ο	ή	λι	ο	ς	μ	ά	λ	ω	ν	α	ν
o	v	o	r	ʝ	ˈa	s	c	o	ˈi	ʎ	o	s	m	ˈa	l	o	n	a	n

A custom software processed the orthographic and phonetic types lists with a greedy assignment algorithm using an expanded list of possible phoneme-grapheme mappings originally based on Petrounias (2002, Table 15.2, pp. 498–502). In the vast majority of cases, simply assigning the longest letter sequence matching the current phoneme resulted in correct parsing (i.e., one that accounted for all phonemes with graphemes in the matching list and accounting for all letters in the orthographic string). A few special cases were identified and treated separately, such as two-phoneme letters (⟨ξ⟩ and ⟨ψ⟩) and context-dependent palatal allophonic variants of [i], without affecting the strictly sequential principle of alignment.

Results

Grapheme-phoneme consistency and entropy

There were 118 unique grapheme-phoneme mappings ("sonographs," in the terminology of Spencer, 2009) accounting for the 147,398,522 (frequency-weighted) phoneme-grapheme pairs in the complete corpus. Additional grapheme-phoneme pairs are possible, phonotactically and orthographically allowed, and may possibly occur in very low frequency or loan words not included in this corpus. Table 1 shows the occurring mappings, grouped and counted by phoneme, and the proportion of occurrence for each grapheme (over the total count of the corresponding phoneme). The proportion of the most frequent grapheme for each phoneme is displayed first, in a separate column to the left of the smaller proportions following it. The token sum of the most frequent grapheme for each phoneme divided by the total number of grapheme-phoneme pairs in the corpus is .803. To the extent that this ratio can be considered to be a single-number estimate of the consistency of phoneme-to-grapheme mapping, Greek is then 80.3% consistent in the feedback (spelling) direction by token count.

Table 2 shows the same mappings, grouped and counted by grapheme, with the corresponding proportions now referring to sums over graphemes. By a similar calculation of an estimate for the consistency of grapheme-to-phoneme map-

ping, Greek is 95.1% consistent in the feedforward (reading) direction. The grapheme unit size was selected as most appropriate because the corresponding calculation using single letters instead of graphemes resulted in a very substantially lower consistency estimate (80.3%) and a greater number of mappings (173) in the reading direction. Table 3 lists these and corresponding estimates derived from type counts. Ignoring the stress diacritic and treating stressed and unstressed letters (and phonemes) as identical would result in 88 grapheme-phoneme mappings with an overall token consistency of 96.0% in the feedforward and 80.8% in the feedback direction.

Entropy values calculated following Borgwaldt et al. (2004, p. 171) are listed in Table 4 under "Type counts." In addition, entropy values under "Token counts" were calculated using the frequency weighted word list as in the calculations of consistency above. Because of the difficulty in defining graphemes in less transparent orthographies, Borgwaldt et al. used only word-initial mappings. Therefore we also list calculations based on word-initial mappings, for comparison. Table 4 also lists entropy for vowels and consonants separately, following Borgwaldt et al. (2005, except that lemma type counts were not available for our corpus, so word form type and token counts were used instead). The distinction between consonants and vowels was always made on the basis of the phonemes (not letters) for each individual mapping, so that a given letter might be counted as a vowel in one case and as a consonant in another (e.g., ⟨υ⟩→[u] vs. ⟨υ⟩→[v]).

Regularity of graphophonemic conversion

Table A1 in the Appendix lists an ordered set of 80 rules (originally based on the set of mappings in Petrounias, 2002, pp. 498–502) that can transcribe correctly the complete text corpus based on the word-form letter sequences only, without any additional information. Because of potential overlap or ambiguity, rules are ranked in fixed order such that more specific rules take precedence over more general ones (nonoverlapping rules, for which rank does not matter, are listed in Greek alphabetical order). Many rules are actually group rules in that each applies over a set of related letters or letter combinations, specified in the Appendix, Table A2. These sets indicate possible (not actual) combinations; full expansion of the group rules using the complete letter sets indicated in Table A2 results in 4163 individual rules, of which only 525 apply at least once in the analyzed corpus. Obligatory preceding and following contexts for the rules are fully indicated in Table A1.

A number of rules are marked as "optional" because of the ambiguous CiV sequences, which can be parsed in two ways. Specifically, because the cases with palatal consonants are more specific to the CiV phenomenon, whereas the [i] case amounts simply to pronouncing each component of the CiV as it would be pronounced in other contexts, the rule set in Table A1 lists the palatal rules as special cases, having precedence over the more general mappings. Rules 8, 10, 23, 24, 29, 30, 41, 42, and 67 correspond to the two-grapheme parsings of the CiV, in which the Ci together map onto a palatal consonant. Rules 17, 73, and 74 correspond to the three-grapheme parsings, in which the "unstressed i" grapheme alone (set U in Tables A1 and A2) maps onto a palatal consonant. These special rules are listed as optional, because the correct pronunciation, being lexically determined, cannot be derived from orthographic or phonological information at the grapheme-phoneme level. In the discussion, the term "CiV rules" refers to (optional) rules that lead to a palatal consonant.

The number of times each rule actually applies and leads to a correct pronunciation, divided by the number of times the rule should apply according to the matching criteria and its rank in the rule set, constitutes an estimate of the regularity for the rule, and is listed in Table A1, computed on token counts. Regularity estimates lower than 1.00 are caused by (a) an optional rule taking precedence, and thus precluding application of a non-optional but lower-ranked alternative, (b) phonologically stressed syllables not orthographically marked with a diacritic due to the monosyllable rule (rules 5, 71, 75, 77, and 79; see the section on stress and monosyllables below), and (c) clear-cut cases of exceptions, such as [mp] and [nt] in recent loans (the latter are quite rare, accounting for .03–.56% of the corresponding digraph rules).

At the word level, regularity can be calculated as the proportion of words read correctly on the basis of their orthography alone. A word is considered correct when all of its phonemes are correctly mapped. When the optional rules are included in the rule set, word-level regularity is 92.7% (by token count). When the optional rules are removed from the rule set, word-level regularity is 95.3%. Finally, when optional rules are allowed to apply optionally, with either outcome counting as correct, the word-level regularity estimate reaches 97.3%. This latter value surely overestimates the regularity of the system because it sidesteps the whole irregularity problem raised by the existence of the "optional" rules. However, it is useful to keep in mind that there is a certain order even in these cases because the ambiguity is always between two well-defined alternatives and not totally unpredictable as it might be in more opaque orthographies.

Stress and monosyllables

It is not possible to distinguish which disyllabic tokens in the corpus bear phonological stress without examining the phrase context of each individual occurence. It is, however, possible, to reach an approximate estimate for the monosyllables, because examination of a sample of occurrences indicates, in the vast majority of cases, whether a stressed or unstressed reading strongly predominates. Therefore, to estimate the effect inconsistencies in application of the orthographic stress diacritic relative to the phonological stress, we classified all monosyllables in the corpus according to whether they bear phonological stress or not (the cumulative frequency of unclassifiable monosyllable types was negligible).

There were 466 monosyllabic types (0.2%) accounting for 37.6% of the total token count (11,108,247 tokens). Of these

Table 1
Phoneme to grapheme mappings, grouped by phoneme and sorted by within-phoneme proportions. Each line refers to a single phoneme-grapheme mapping in the corpus. F: Relative frequency (percent occurrences of all phoneme tokens) of this phoneme in the corpus. Pair proportion: Percentage of this phoneme-grapheme pair as a proportion of all occurrences (tokens) of this phoneme. The proportion of the dominant mapping is listed first, on the the left; other mappings follow, on the right. A minimum of one significant digit is shown.

			Pair proportion (%)	
Phoneme	F (%)	Grapheme	Highest	Other
a	8.26	α	100.0	
b	.18	μπ	100.0	
c	1.96	κ	97.2	
		κι		2.6
		κκ		.17
		κυ		.02
		κει		.0006
d	.56	ντ	100.0	
ð	1.87	δ	100.0	
ɛ	6.71	ε	78.0	
		αι		22.0
f	1.35	φ	66.5	
		υ		28.6
		ύ		4.8
		υφ		.06
		ύφ		.01
g	.08	γκ	85.5	
		γγ		14.5
ɟ	.07	γκ	57.9	
		γγ		41.6
		γκι		.38
		γγι		.09
ŋ	.03	γ	95.1	
		ν		4.9
i	10.77	η	39.1	
		ι		33.6
		ει		10.8
		υ		10.7
		οι		5.3
		ϊ		.41
		ϋ		.09
		υι		.02
ʝ	.72	γ	71.5	
		ι		22.1
		γι		6.1
		ει		.17
		γυ		.11
		υ		.04
		γει		.01
ɣ	.78	γ	99.9	
		γγ		.07
k	2.16	κ	99.6	
		κκ		.38
ks	.47	ξ	99.1	
l	2.43	λ	86.7	
		λλ		13.3
ʎ	.04	λι	69.6	
		λει		23.1
		λλι		7.3
m	3.34	μ	96.8	
		μμ		3.0
		μπ		.19
n	6.25	ν	99.8	
		νν		.23
ɲ	.05	νι	72.9	
		ι		16.6
		υ		3.7
		οι		3.6
		νοι		1.6
		ννι		1.5
		νει		.11
o	6.87	ο	76.7	
		ω		23.3
p	4.22	π	99.9	
		ππ		.05
ps	.16	ψ	100.0	
ɾ	4.59	ρ	99.5	
		ρρ		.48
s	8.34	σ	54.6	
		ς		44.8
		σσ		.57
ts	.04	τσ	89.4	
		τς		10.6
t	8.43	τ	99.9	
		ττ		.14
θ	1.27	θ	100.0	
u	2.14	ου	100.0	
v	.82	β	77.5	
		υ		14.5
		ύ		7.9
		ββ		.03
		υβ		.004
x	.74	χ	100.0	
ç	.67	χ	85.0	
		οι		7.0
		ι		6.9
		ει		.60
		χι		.39
		χει		.09
		υ		.08
		χυ		.001
z	.59	ζ	57.1	
		σ		42.9
ʣ	.01	τζ	92.9	
		ντζ		7.1
ˈa	2.21	ά	95.2	
		α		4.8
ˈɛ	2.60	έ	79.6	
		ε		12.8
		αί		7.4
		αι		.15
ˈi	4.40	ή	33.7	
		ί		29.7
		εί		20.1
		ύ		9.9
		οί		4.3
		ι		.91
		η		.70
		ει		.57
		ΐ		.05
		ΰ		.0005
		υί		.0003
ˈo	3.29	ό	72.7	
		ώ		24.7
		ω		1.5
		ο		1.0
ˈu	.54	ού	99.4	
		ου		.58

Note: Phoneme pairs [ks] and [ps] appear under "Phoneme" because they map to single letters ⟨ξ⟩ and ⟨ψ⟩, respectively. This results in slight overestimation of consistency for [k], [p], and [s].

Table 2
Grapheme to phoneme mappings, grouped by grapheme and sorted by within-grapheme proportions. Each line refers to a single grapheme-phoneme mapping. F: Relative frequency (percent occurrences of all grapheme tokens) of this grapheme in the corpus. Pair proportion: Percentage of this grapheme-phoneme pair as a proportion of all occurrences (tokens) of this grapheme. The proportion of the dominant mapping is listed first, on the the left; other mappings follow, on the right. A minimum of one significant digit is shown.

			Pair proportion (%)	
Grapheme	F (%)	Phoneme	Highest	Other
α	8.37	a	98.7	
		ˈa		1.3
ά	2.11	ˈa	100.0	
αι	1.48	ɛ	99.7	
		ˈɛ		.3
αί	.19	ˈɛ	100.0	
β	.64	v	100.0	
ββ	.00	v	100.0	
γ	1.33	ɣ	58.7	
		ʝ		38.9
		ŋ		2.4
γγ	.04	ɟ	72.3	
		g		26.5
		ɣ		1.2
γγι	.00	ɟ	100.0	
γει	.00	ʝ	100.0	
γι	.04	ʝ	100.0	
γκ	.11	g	60.8	
		ɟ		39.2
γκι	.00	ɟ	100.0	
γυ	.00	ʝ	100.0	
δ	1.87	ð	100.0	
ε	5.56	ɛ	94.0	
		ˈɛ		6.0
έ	2.07	ˈɛ	100.0	
ει	1.19	i	97.4	
		ˈi		2.1
		ç		.3
		ʝ		.1
εί	.89	ˈi	100.0	
ζ	.34	z	100.0	
η	4.24	i	99.3	
		ˈi		.7
ή	1.48	ˈi	100.0	
θ	1.27	θ	100.0	
ι	3.87	i	93.4	
		ʝ		4.1
		ç		1.2
		ˈi		1.0
		ɲ		.2
ί	1.31	ˈi	100.0	
ϊ	.04	i	100.0	
ΐ	.00	ˈi	100.0	
κ	4.05	k	53.0	
		c		47.0
κει	.00	c	100.0	
κι	.05	c	100.0	
κκ	.01	k	70.9	
		c		29.1
κυ	.00	c	100.0	
λ	2.10	l	100.0	
λει	.01	ʎ	100.0	
λι	.03	ʎ	100.0	
λλ	.32	l	100.0	
λλι	.00	ʎ	100.0	
μ	3.24	m	100.0	
μμ	.10	m	100.0	
μπ	.18	b	96.6	
		m		3.4

			Pair proportion (%)	
Grapheme	F (%)	Phoneme	Highest	Other
ν	6.23	n	100.0	
		ŋ		.03
νει	.00	ɲ	100.0	
νι	.04	ɲ	100.0	
νν	.01	n	100.0	
ννι	.00	ɲ	100.0	
νοι	.00	ɲ	100.0	
ντ	.56	d	100.0	
ντζ	.00	ʤ	100.0	
ξ	.47	ks	100.0	
ο	5.30	o	99.4	
		ˈo		.6
ό	2.39	ˈo	100.0	
οι	.62	i	92.1	
		ç		7.5
		ɲ		.3
οί	.19	ˈi	100.0	
ου	2.14	u	99.9	
		ˈu		.1
ού	.53	ˈu	100.0	
π	4.22	p	100.0	
ππ	.00	p	100.0	
ρ	4.56	ɾ	100.0	
ρρ	.02	ɾ	100.0	
σ	4.81	s	94.7	
		z		5.3
ς	3.74	s	100.0	
σσ	.05	s	100.0	
τ	8.41	t	100.0	
τζ	.01	ʤ	100.0	
τσ	.04	ʦ	100.0	
τς	.00	ʦ	100.0	
ττ	.01	t	100.0	
υ	1.67	i	69.4	
		f		23.3
		v		7.2
		ɲ		.1
		ç		.03
		ʝ		.02
ύ	.57	ˈi	76.9	
		f		11.6
		v		11.5
ϋ	.01	i	100.0	
ΰ	.00	ˈi	100.0	
υβ	.00	v	100.0	
υι	.00	i	100.0	
υί	.00	ˈi	100.0	
υφ	.00	f	100.0	
ύφ	.00	f	100.0	
φ	.90	f	100.0	
χ	1.31	x	56.4	
		ç		43.6
χει	.00	ç	100.0	
χι	.00	ç	100.0	
χυ	.00	ç	100.0	
ψ	.16	ps	100.0	
ω	1.65	o	97.0	
		ˈo		3.0
ώ	.81	ˈo	100.0	

Note: Phoneme pairs [ks] and [ps] appear under "Phoneme" because single letters ⟨ξ⟩ and ⟨ψ⟩, respectively, map to them.

Table 3
Statistics related to the transparency of Greek orthography. Mean pairs: average number of mappings from a single source unit (letter, grapheme, or phoneme). Total consistency, in percent. V:C is the ratio of vowel to consonant phonemes. Type-Token and Entropy-Consistency: Correlation coefficients calculated as Spearman's ρ between counts/estimates for the corresponding units of each mapping (N=Total pairs). Gra/Phoneme: unique grapheme-phoneme pairs.

					Type counts			Token counts		
Mapping from	to	Total pairs	Mean pairs	Type-Token	Total consist.	Entropy-Consist.	V:C	Total consist.	Entropy-Consist.	V:C
Grapheme	Phoneme	118	1.40	.93	95.7	−.991	.873	95.1	−.986	.915
Letter	Gra/Phoneme	173	4.81	.91	82.5	−.931	.930	80.3	−.933	1.019
1st letter	Gra/Phoneme	64	2.06	.81	93.8	−.963	.567	90.9	−.971	.481
1st letter	Phoneme	54	1.74	.78	93.8	−.954	.567	91.3	−.975	.481
Phoneme	Grapheme	118	3.03	.93	82.9	−.824	.873	80.3	−.810	.915
1st phoneme	Grapheme	64	1.78	.85	93.3	−.913	.567	93.8	−.948	.481
1st phoneme	Letter	54	1.46	.78	93.5	−.976	.567	94.3	−.992	.481

Note: There are fewer mappings from 1st letter to phoneme than from from 1st letter to grapheme-phoneme pairs, because the same letter may map onto the same phoneme as a member of different graphemes.

Table 4
Entropy values for the Greek othography, in both mapping directions, for all entire words and for word-initial units only. Calculations were done both on the type counts (unique word forms) and the token counts (frequency weighted). Gra/Phoneme: unique grapheme-phoneme pairs.

Mapping		Type counts			Token counts		
from	to	Total	Vowels	Consonants	Total	Vowels	Consonants
Grapheme	Phoneme	.163	.033	.198	.167	.085	.177
Letter	Gra/Phoneme	.786	.817	.643	.801	.977	.515
1st letter	Gra/Phoneme	.290	.271	.301	.330	.520	.239
1st letter	Phoneme	.275	.265	.282	.308	.513	.209
Phoneme	Grapheme	.589	.886	.330	.645	1.010	.311
1st phoneme	Grapheme	.277	.406	.203	.251	.626	.071
1st phoneme	Letter	.262	.400	.184	.229	.619	.041

466 word forms, we identified 352 (79.6%) bearing phonological stress, including 94 native Greek words, 189 recent loans (mainly from English, such as "bar," "goal" etc.), 35 exclamatory and onomatopoetic items, some stress-bearing function words, and more than 50 abbreviations either retaining the stressed syllable of the original or having been lexicalized. The total token count of these types was 743,694, i.e., 6.7% of the total monosyllable token count and 2.5% of the corpus. There were also 90 types not bearing phonological stress, primarily function words but also including some abbreviations retaining an unstressed syllable; these accounted for 10,341,231 tokens or 93.3% of the monosyllabic tokens and 35.0% of the corpus. Therefore, as far as monosyllables are concerned, the lack of diacritic demanded by spelling convention coincides with a lack of phonological stress at an estimated rate of more than 90%. Thus, in the reading direction, with respect to stress, the diacritic does not seem to dramatically affect the overall consistency estimate, as expected from the small effect on regularity for the vowel rules noted above.

Discussion

Greek orthography is sufficiently consistent that we have been able to segment and analyze a complete corpus sequentially into graphemes and their relation to phonemes. Our analyses indicate that the grapheme level is appropriate for expressing the widely perceived notion that Greek is simpler to read than to write, because it is at this level that consistency and entropy estimates come out strongly asymmetric in favor of the feedforward direction. Specifically, in the feedforward direction, consistency estimates for individual graphemes range from 53.0% for ⟨ϰ⟩ and 56.4% for ⟨χ⟩ to 100.0% for the majority of graphemes. In the feedback direction, consistency estimates for individual phonemes range from 39.1% for ⟨i⟩ (33.7% for stressed ⟨i⟩) to 100.0% for a small minority of phonemes. A non-parametric comparison of the 84 grapheme consistency estimates to the 39 phoneme consistency estimates confirms that the asymmetry is significant (Mann-Whitney U=761.5, Z=-5.33, asymptotic two-tailed $p < .0005$; the same result is obtained if consistency estimates are weighted by unit token frequency: U=990.0, Z=-3.52, $p < .0005$). It was also clear in the overall counts, where spelling inconsistency (proportion of nondominant

mappings) at the grapheme-phoneme level (19.7%) was four times as large as reading inconsistency (4.9%).

A factor frequently overlooked in orthographic analyses and in reading models concerns suprasegmental features and their orthographic notation, namely stress and metrical patterns and corresponding diacritics that signify them. We have confirmed that the Greek stress diacritic constitutes an additional source of orthographic information, contributing somewhat to overall inconsistency at the metrical level, due to the peculiarities of the spelling rules. Further research in stress-assigning languages should take stress and the diacritics into account, aiming for a full analysis of the orthographic system at multiple levels. Here, the presence of the diacritic does not reduce inconsistency at the grapheme-phoneme level, because ignoring it results in fewer grapheme-phoneme pairs and because its disambiguating role becomes apparent in grapheme segmentation, not in mapping to and from phonemes. Therefore, any beneficial effects of the diacritic would be discernible only in predictability (or regularity) and not in consistency.

In the following, we compare our findings to those from different orthographies and methodologies, discussing the advantages and limitations of different approaches. In addition, we consider two special issues of particular importance, namely the consistency-predictability distinction and the status of the CiV ambiguity.

Cross-linguistic comparisons

The regularity of graphophonemic conversion at the word level (95.3%, excluding the "optional" rules) was found to be higher for Greek than for German (90.4%) or English (79.3%) as reported by Ziegler et al. (2000). However, these numbers may not be directly comparable, because German and English estimates were derived on the basis of monosyllables only.

The overall consistency for Greek at the grapheme-phoneme level was 95.1% in the feedforward and 80.3% in the feedback direction. There are few directly comparable estimates from other languages. According to Spencer (2007, p. 306), Hanna, Hanna, and Hodges (1966) reported a feedback consistency of 73% at the phoneme-grapheme level, corresponding to 50% at the word level. However, summation of the most frequent sonograph probabilities for each of the 163 graphemes and for each of the 39 phonemes in the "adult 3K" corpus of Spencer (2009, supplemental material, Appendix C) would result in consistency estimates of 77.6% and 57.0% for the feedforward and feedback directions, respectively, at the grapheme-phoneme level.

Ziegler et al. (1996, 1997) examined only spelling bodies (rimes) of monosyllables and reported their bottom-line findings in terms of proportion of words that admit more than a single mapping. They considered the presence of any alternatives as indicative of inconsistency regardless of the relative proportion of alternatives, and they did not distinguish dominant from other mappings. Therefore, their estimates for French (79.1% spelling-inconsistent and 12.4% reading-inconsistent) and English (72.3% and 30.7%, respectively) may be gross overestimates of the overall inconsistency of those orthographic systems.

On the basis of the entropy results for word-initial letter type counts, closely matching the methodology of Borgwaldt et al. (2004), Greek is about equally ambiguous in both the feedforward (1st letter → 1st phoneme, H = .275) and the feedback (1st phoneme → 1st letter, H = .262) direction, being similar to Dutch in reading and similar to French in writing (comparing to Figure 1 of Borgwaldt et al., 2004, p. 175). For reading, Greek is less transparent than Hungarian and more transparent than French, German, and English. For spelling, Greek is less transparent than Hungarian, Dutch, and German, and more transparent than English.

Considering entropy calculations for vowels and consonants separately (and comparing to Figure 2 of Borgwaldt et al., 2005, p. 219), for the feedforward (reading) direction only, Greek is less transparent than Hungarian and Italian and more transparent than Portuguese, Dutch, French, German and English, as far as vowels are concerned. With respect to consonants, Greek is only more transparent than French, exceeding in consonant entropy all these other languages. However, these cross-language comparisons are valid only to the extent that word-initial mappings are representative of the full words in every language.

Representativeness of analyzed words or word parts

Previous studies of orthographic transparency have usually analyzed a small subset of all mappings that occur in the written language. A common restriction seen in the literature is to analyze monosyllables only (Ziegler et al., 1996, 1997, 2000). Our results suggest that although this approach may be justified for some languages such as English, it is not appropriate for cross-linguistic comparisons.

Specifically, Greek has very few monosyllabic word forms and the majority of them is not representative of the language. Most are function words, whereas others are recent loans and abbreviations, strongly atypical in their syllabic structure. Therefore no analysis based on monosyllables only could be expected to yield an outcome representative for the orthography of the language as a whole. In English monosyllables are sufficiently numerous and diverse to permit a meaningful analysis. However, it remains to be systematically investigated whether findings from monosyllables can be generalized, due to possible differences from polysyllables (Kessler & Treiman, 2001; Lété et al., 2008). In the context of cross-linguistic investigations, with some languages containing predominantly multisyllabic words, it is important to extend analyses to more inclusive and representative samples of the vocabulary.

To overcome the monosyllable restriction and perform meaningful cross-linguistic comparisons, Borgwaldt et al. (2004, 2005) analyzed word-initial mappings using single letters. For comparison, we have calculated entropy on both word-initial only as well as full-word mappings (Table 4). The results show that, in the feedforward (reading) direction, using word-initial single-letter mapping to

phonemes results in entropy values almost three times less than values calculated over whole-word letter-to-phoneme mappings and about twice as large as values calculated over whole-word grapheme-to-phoneme mappings. In the feedback (spelling) direction, using word-initial mappings results in entropy values less than half of the values calculated over whole-word mappings, regardless of the target units (letters or graphemes). Therefore, using word-initial letter mappings in Greek results in a gross underestimation of orthographic transparency compared to calculations over letters using the entire words. This may be attributed to the nonrepresentativeness of word-initial mappings in failing to exhibit the full spectrum of mapping complexities that may be encountered in different parts of the words.

Breaking down the entropy analysis into separate calculations for vowels and consonants, we can compare the effect of restricting our analysis to word-initial letters on the entropy estimates for each type of phoneme. Focusing on the left side of Table 4 (type counts, to match the methodology of Borgwaldt et al. as closely as possible), we see that, in the feedforward direction, the entropy of word-initial vowels is somewhat lower than that of word-initial consonants, whereas the reverse is true when full-word letter mappings are considered. In the feedback direction, there are differences in the relative values of vowels and consonants depending on whether word-initial or full-word mappings are taken into account, but the direction of the pairwise comparison remains unaffected. However, the relationship between vowel and consonant entropy seems substantially distorted, in both directions, in comparison with phoneme/grapheme mappings.

Our results indicate that the restriction to word-initial mappings may be vulnerable to systematic distortions due to differences in the distribution of phonemes (and letters/graphemes) across word positions. In Greek, this may be due to the relative proportion of vowels to consonants being much lower (about half) word-initially than it is overall (see Table 3, columns "V:C"), perhaps reflecting the preponderance of CV syllables in the language. Word-initial mappings will be representative only insofar as the distribution of syllable types in the language is uniform, that is, syllables with consonantal onsets are as frequent as syllables without onsets, and to the extent that onset and coda phonotactics and spelling patterns are similar.

In languages with rime-level consistency, such as English, onset spelling patterns are not representative of all spelling patterns. Indeed, some analyses of orthographic transparency in English have focused on rimes only, ignoring onsets, because the rime is where most inconsistencies seem to be encountered. In languages with a predominance of CV syllables, such as Greek, onset spelling patterns will be nonrepresentative in containing a larger proportion of consonants. If consonant mappings are not as consistent as vowel mappings, then consistency estimates based on word-initial mappings only are unlikely to constitute valid indices for the language. Word-initial mappings may additionally undersample the graphophonemic mapping space if certain graphemes or phonemes cannot occur syllable-initially. Our calculations suggest that both of these conditions are present in Greek and thus the validity of using only word-initial mappings is questionable at best.

CiV and regularity

Whether or not the origin of the CiV inconsistency can be diachronically traced to a distinction between vernacular and literary vocabulary (Petrounias, 2002), it does not help determine what is "regular" and what is "irregular" behavior. The mapping (C, i, V) may be regular because it is more general; or the (Ci, V) mapping may be regular because in the ordered rule set it must constitute a higher-precedence rule. Further research will be needed to determine whether one of the alternatives can be considered to be a rule, to which the other alternative would constitute an exception.

In the linguistic sense, a "rule" may correspond to the "default" or "unmarked" behavior, that is, what happens when nothing special applies. There is no quantitative implication for the regularity corresponding to this rule. So, the regular German plural concerns a minority of nouns (Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1995) whereas the regular English past tense concerns a majority of verbs (Pinker & Prince, 1988). In contrast, in the reading literature, "rules" tend to be interpreted as accounting for maximum regularity, thus English regular readings correspond to the most frequent grapheme-phoneme mappings (Coltheart et al., 2001). The former sense is theoretical and can be investigated empirically with novel or otherwise unmarked stimuli. The latter sense, however, is distributional and can be examined by reference to the corpus data, by comparing the frequency of application vs. non-application of rules in question.

As shown in Table A1, all CiV rules, with the notable exception of rule 8 involving ⟨γ⟩, apply less often than not. This suggests that the palatal pronunciation may be exceptional and the [i] option regular. The great variability of observed rates of application and the nonconformance of ⟨γ⟩ to this pattern limit the confidence of this assertion. Further analyses were undertaken by breaking down the group rules into their constituent sets, by i-grapheme and by vowel; and further breaking down by stress (stressed vs. unstressed V) and by position in the word (initial, medial, final syllable). In each of these groupings, the CiV rules apply less often than not in the great majority but not all cases. Therefore, from an overall frequency perspective, the palatal readings of the CiV clusters appear to constitute the exception rather than the rule. However, the situation is not so clear if we take into account the many different consonant-i-vowel combinations that are possible.

The letter sets indicated in Table A2 cover all possible letter combinations whether or not they are ever encountered in any word spellings. In fact, of the 3744 individual CiV rules that can be obtained by fully expanding the 12 group rules (rows 8, 10, 17, 23, 24, 29, 30, 41, 42, 67, 73, and 74 in Table A1), only 650 letter combinations ("cases") appear in the corpus. Of these, in only 298 cases does the corresponding individual CiV rule actually apply at least once. The CiV

rule applies more often than not in 137 of these 298 cases, but the sum token count of actual rule applications in cases in which the CiV rule dominates (i.e., applies more often than not) is approximately equal to the total count in cases in which the rule appears to recede (50.2% vs. 49.8%). This happens because the individual CiV rules dominate in fewer but more frequent cases. Therefore, this corpus analysis cannot provide a clear answer to the question whether the CiV rules correspond to the "regular" case in Greek grapheme-to-phoneme mapping or, indeed, if a case for a regular/irregular distinction can be made regarding this phenomenon.

Effects of boundaries and context on consistency

Analyses of consistency are necessarily limited to the discrete units used in the computations, be they letters, graphemes, or larger units. However, neighboring units may affect the distribution of the mappings. Moreover, with the exception of single letters, it is not always clear what constitutes a unit, that is, where the segmentation boundaries lie. These issues are not captured by the consistency estimates. However, they may influence the effects of consistency on reading and spelling performance, if context is taken into account for processing at the basic phoneme-grapheme level (Kessler & Treiman, 2001).

Mapping consistency is a notion distinct from predictability, because context may determine (or bias) the appropriate mapping among a set of alternatives. For example, [z] can be spelled with ⟨ζ⟩ (57.1%) or ⟨σ⟩ (42.9%), so its consistency is only 57.1%. However, the correct spelling depends on the context: when a voiced consonant follows, it is ⟨σ⟩. Therefore, spelling of [z] is 100% predictable. In this case, the low consistency of the phoneme is misleading in that the phonological context provides all necessary information for spelling. Conversely, the grapheme ⟨γι⟩ is always pronounced [ʝ], so its consistency is 100%. However, it is not possible to determine, from the letter sequence alone (even taking context into account), whether the pair of letters ⟨γι⟩ constitute a grapheme or, rather, should be parsed as two graphemes, ⟨γ⟩ and ⟨ι⟩, to be pronounced [ʝ] and [i], respectively. In this case the high consistency of the grapheme belies the unpredictability of its status as a grapheme.

In the case of mapping phonemes to graphemes, the predictability of seemingly inconsistent phonemes can be estimated by examining the contexts of each phoneme in which different graphemes appear. There are no known constraints at the phonological level to the alternative spellings of vowels [i], [e], and [o], which are the major sources of inconsistency. In contrast, spelling of [z] is predictable, as noted above; as is spelling of [s] (⟨ς⟩ at word end, ⟨σ⟩ otherwise). Spellings for [v] and [f] are constrained in that the ⟨υ⟩ variants may only appear after ⟨α⟩, ⟨ε⟩, or ⟨η⟩, though not obligatorily (cf. ˈklefti→κλέφτη, ˈpsefti→ψεύτη). The palatal consonants are also constrained. Spellings of [ʝ], [ç], [c], [ɟ] with a single consonant letter appear only before a high vowel (ˈcita→κοίτα,ˈçɛɾi→χέρι), otherwise an i-grapheme is added (ˈcali→κυάλι,ˈçoni→χιόνι). When preceded by a same-voicing consonant, [ʝ] and [ç] may be spelled with an i-grapheme only (ˈtɛtçɛs→τέτοιες, ˈisçɛs→ίσιες, ˈvʝos→βιός; but cf. ˈɛsçɛs→έσχες, ˈvʝɛs→βγες). Likewise, when preceded by [m], [ɲ] is spelled with an i-grapheme only (ziˈmɲa→ζημιά; but there are exceptions: liˈmɲa→Λημνιά).

By adding the phoneme tokens for [z], [ts], and [s], for which the correct spelling is entirely predictable, to the sum of consistent mappings, we can derive an estimate of the *lower bound* of predictability for Greek feedback mapping (spelling) at about 84.3%. Is it useful to take context-based predictability into account when considering effects of consistency on spelling performance? If consistency alone determines spelling difficulty then [z], [ts], [s], and the palatal consonants should be spelled incorrectly as frequently as the vowels [i], [e], and [o], which are comparably inconsistent. Although specific estimates of these spelling errors are not yet available for Greek, the fact the vowel spellings are frequently studied but consonant spellings are not suggests that grapheme-level consistency does not suffice and context or multiple-size units may have to be taken into account in the quantification of transparency that is relevant for psycholinguistic investigation.

Grain size of transparency analysis

Notwithstanding the occasional effects of context, our results suggest that the grapheme seems to be the appropriate-size unit for analyzing the transparency of Greek orthography. On the one hand, the smaller-size letter units are less consistent and therefore unable to adequately capture the systematicity in graphophonemic mappings. On the other hand, no larger sublexical units, such as rimes, seem to be necessary or useful, because resorting to larger units would increase granularity without improving consistency. Larger unit sizes would not help resolve the predominant source of inconsistency in the feedforward direction, namely the CiV cases, because those are determined lexically. The rime-body level of analyzing English monosyllabic words is not appropriate for other languages with many polysyllabic words as well: French pronunciation and spelling ambiguities are not reduced when rime-body mappings are used instead of grapheme-phoneme mappings (Lété et al., 2008; Peereman & Content, n.d.).

The calculation of quantitative estimates of orthographic transparency is not an end in itself. It is meaningful to the extent that it can contribute to the generation or testing of specific psycholinguistic hypothesis, or to the selection, control and construction of proper experimental stimuli, for work both within and between languages. An important issue concerns the psychological reality of different grain sizes across languages. Analyses are often performed to minimize ambiguity. However, there is no a priori reason that maximizing systematicity is a valid objective or that it results in estimates that are relevant for modeling. A more solid foundation can be sought in correlations between transparency analyses and performance on reading tasks. If readers rely on a particular orthographic and phonological grain size when they read (or spell), then transparency measures based on the corresponding functional units will correlate more highly with perfor-

mance on the reading (or spelling) tasks.

Along these lines, Treiman et al. (1995) showed that word naming was affected not only by the consistency of individual graphemes but also by the consistency of reliable units such as consonantal onsets and orthographic rimes. Performance was not affected by the consistency of less reliable combinations such as the head (initial consonant and vowel). Such findings seem to establish the psycholinguistic relevance of onset/rime units in English, where large ambiguities exist at the phoneme/grapheme level. More recently, Borgwaldt et al. (2005) found that higher onset letter entropies (i.e., less consistent word-initial letter-phoneme mappings) were associated with longer naming latencies in Italian, Dutch, and English, three languages differing greatly in orthographic transparency, and argued that letters are important functional units that should not be ignored in favor of larger grain sizes. To make further progress on this issue, additional research is needed, examining contrasting predictions from graphemes and other units of various sizes and word positions, in both directions, across a variety of orthographies.

A comparison of entropy (or consistency) values calculated over graphemes vs. letters (first two rows of Tables 4 and 3) provides a clue towards resolving the unit issue for Greek. Calculations using single letters result in very high estimates of feedforward ambiguity—higher, in fact, than the corresponding estimates for the feedback (spelling) direction. This contradicts the commonly held notion that Greek is consistent for reading and inconsistent for spelling. However, intuitive notions may well be incorrect. More importantly, these estimates seem to run counter to available cross-linguistic data on the relative transparency of Greek, such as the findings of Seymour et al. (2003), according to which Greek is placed near the top of the list of transparent orthographies, based on the accuracy and speed of reading simple words and nonword by beginning readers (Grade 1). Additional evidence consistent with the notion that Greek is feedforward consistent and that the grapheme/phoneme is the appropriate level of analysis (Goswami, Porpodas, & Wheelwright, 1997; Porpodas, Pantelis, & Hantziou, 1990; Porpodas, 1999) has been reviewed by Ziegler and Goswami (2005). A more stringent test to determine the appropriate level of analysis can be derived from our entropy data (Table 4). Specifically, in the feedforward direction, vowels are more ambiguous than consonants at the letter level, but less ambiguous at the grapheme level, and the difference is quite substantial in both cases. Insofar as mapping consistency affects reading efficiency, we should expect vowels to affect reading performance more than consonants, if the input is analyzed at the single letter level, or less, if the input is analyzed by graphemes. Due to the rapid learning of letter combinations, such effects may only be discernible at the earliest stages of learning to read, if at all.

In the preceding discussion we have assumed that there is a single most appropriate level of analysis, psychologically real and dominating sublexical reading performance. This view is consistent with the DRC approach (Coltheart et al., 2001), in which only grapheme/phoneme mappings are hypothesized to exist in the nonlexical route, excluding larger-size units (see also Coltheart, Curtis, Atkins, & Haller, 1993, p. 603, for empirical justification). However, it remains plausible that more than a single level of cross-code representation is psychologically real, consistent with the grain-size theory (Ziegler & Goswami, 2005), which posits a developmental progression through multiple levels of representation, from coarser to finer, supported by the distributional properties of the developing lexicon, such that denser lexical neighborhoods facilitate finer phonological analysis. In a similar vein, Ehri (2005) describes the development of sight-word vocabulary as a process of forming connections between not only graphemes and phonemes but also between orthographic and phonological sublexical representations of various sizes. At the final phase of this progression, readers familiar with the alphabetic system of their language retain entire words in memory. However, the importance of larger-sized units may be smaller in more transparent orthographies.

Comparisons among indices and counts

Alternative modes of quantifying transparency, based on diverging functional hypotheses, may bear substantially on the resolution of theoretical matters. For example, the notion of regularity, critically hinging on the existence of rules, is distinct from the notion of consistency, which requires a fixed set of mappings among appropriate-sized units. The DRC simulations of Coltheart et al. (2001) did not support the hypothesis that consistency may arise as an effect lexical neighborhoods. They interpreted consistency effects as caused by the serial application of overlapping graphophonemic rules, when the correct rule becomes activated after an inappropriate one. In contrast, Zevin and Seidenberg (2006) reviewed graded consistency effects in word and nonword reading and attributed them to the statistical properties of spelling-sound mappings. They simulated previous behavioral findings with a parallel distributed (PDP) model of reading aloud and concluded against graphophonemic mapping rules that are not sensitive to their probability of application. In Greek, contrasting predictions regarding regularity and consistency may be derived from the CiV pattern and the context effects on consistency.

The notion of entropy has been introduced as a more comprehensive measure of consistency than the percentage of dominant mappings, because it takes into account the relative proportions of nondominant mappings, and not as a theoretically distinct construct. It remains to be empirically demonstrated whether entropy values are more psycholinguistically relevant than consistency percentages. However, it may not be simple to disentangle the two if they are too similar to lead to differential predictions. For Greek, the correlations between consistency percentages and entropy values are listed in Table 3 (columns "Entropy-Consistency"). The two indices correlate very highly, especially in the feedforward direction. There is some divergence in the feedback direction, suggesting that situations like the [ɡ]/[ç] example mentioned in the introduction may allow differential predictions regarding the ease of initial learning to spell. With proper control

of the order and amount of teaching and practice for each mapping, the most appropriate consistency index may be determined.

On the issue of type vs. token counts, our analyses do not seems to offer directions for resolution. Even though there are some differences in the results using the two kinds of counts, comparisons preserve their direction and, usually, their relative proportions, both for entropy (Table 4) and for consistency percentages (Table 3). In the feedforward direction, using type or token counts does not seem to affect total entropy or consistency values very much. In the feedback direction, using type counts leads to higher total entropy estimates on whole-word mappings but lower on word-initial mappings, compared to using token counts. The relative difference in entropy values follows the correlation between type and token counts (listed in Table 3, column "Type-Token"). Using types instead of tokens has the largest effect for mappings between word-initial phonemes and letters, that is, in the condition chosen by Borgwaldt et al. (2004) for their calculations. However, as noted above, this condition does not seem to be the most appropriate one. The choice between type and token counts, then, remains to be made on the basis of psycholinguistic evidence and theory of the sort discussed by Conrad et al. (2008).

Conclusion

In this work, we have provided quantitative indices of orthographic transparency for Greek. We have compared our findings with similar data reported for other orthographies and we have discussed limitations and implications arising from particular methodological choices and shortcuts previously applied. Our results indicate that restricting the analysis to unrepresentative samples of the orthography, such as monosyllabic words or word-initial letters, may distort the outcomes and render cross-linguistic comparisons uninterpretable. However, it remains to be established whether meaningful cross-linguistic comparison on a common level of analysis is possible. If the statistical properties of each orthography determine the psychologically dominant units of processing then transparency analyses may help identify that unit but will not be amenable to direct comparisons. The critical cross-linguistic work, then, must take place at the theoretical level of representations and processes, along the lines of dual-route or distributed models of word recognition and reading aloud.

References

Balota, D. A., Yap, M. Y., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 285–375). London: Academic Press. doi:10.1016/B978-012369374-7/50010-9

Borgwaldt, S. R., Hellwig, F. M., & De Groot, A. M. B. (2004). Word-initial entropy in five languages: Letter to sound and sound to letter. *Written Language and Literacy*, *7*, 165–184. doi:10.1075/wll.7.2.03bor

Borgwaldt, S. R., Hellwig, F. M., & De Groot, A. M. B. (2005). Onset entropy matters – Letter-to-phoneme mappings in seven languages. *Reading and Writing*, *18*, 211–229. doi:10.1007/s11145-005-3001-9

Burani, C., Barca, L., & Ellis, A. W. (2006). Orthographic complexity and word naming in Italian: Some words are more transparent than others. *Psychonomic Bulletin and Review*, *13*, 346–352.

Burt, J. S., & Blackwell, P. (2008). Sound-spelling consistency in adults' orthographic learning. *Journal of Research in Reading*, *31*, 77–96. doi:10.1111/j.1467-9817.2007.00362.x

Chalamandaris, A., Raptis, S., & Tsiakoulis, P. (2005, September). Rule-based grapheme-to-phoneme method for the Greek. In *Interspeech 2005: 9th European conference on speech communication and technology* (pp. 2937–2940). Lisbon, Portugal.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151–216). London: Academic Press.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processsing approaches. *Psychological Review*, *100*, 589–608. doi:10.1037/0033-295X.100.4.589

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. doi:10.1037/0033-295X.108.1.204

Conrad, M., Carreiras, M., & Jacobs, A. M. (2008). Contrasting effects of token and type syllable frequency in lexical decision. *Language and Cognitive Processes*, *23*, 296–326. doi:10.1080/01690960701571570

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, *9*, 167–188. doi:10.1207/s1532799xssr0902_4

Goswami, U., Porpodas, C., & Wheelwright, S. (1997). Children's orthographic representations in English and Greek. *European Journal of Psychology of Education*, *3*, 273–292.

Grainger, J., & Ziegler, J. C. (2008). Cross-code consistency in a functional architecture for word recognition. In E. L. Grigorenko & A. J. Naples (Eds.), *Single-word reading: Behavioral and biological perspectives* (pp. 129–158). New York: Erlbaum.

Hanna, P. R., Hanna, J. S., & Hodges, R. E. (1966). *Phoneme-grapheme correspondences as cues to spelling improvement*. Washington: U.S. 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.

Hofmann, M. J., Stenneken, P., Conrad, M., & Jacobs, A. M. (2007). Sublexical frequency measures for orthographic and phonological units in German. *Behavior Research Methods*, *39*, 620–629.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language*, *46*, 723–750. doi:10.1006/jmla.2001.2827

Kessler, B., & Treiman, R. (2001). Relationships between sounds and letters in English monosyllables. *Journal of Memory and Language*, *44*, 592–617. doi:10.1006/jmla.2000.2745

Kessler, B., Treiman, R., & Mullennix, J. (2008). 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). New York: Erlbaum.

Lété, B., Peereman, R., & Fayol, M. (2008). Consistency and

word-frequency effects on spelling among first- to fifth-grade French children: A regression-based study. *Journal of Memory and Language*, *58*, 952–977. doi:10.1016/j.jml.2008.01.001

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, *29*, 189–256. doi:10.1006/cogp.1995.1015

McKague, M., Davis, C., Pratt, C., & Johnston, M. B. (2008). The role of feedback from phonology to orthography in orthographic learning: An extension of item-based accounts. *Journal of Research in Reading*, *31*, 55-76. doi:10.1111/j.1467-9817.2007.00361.x

Moscoso del Prado Martín, F., Ernestus, M., & Baayen, R. H. (2004). Do type and token effects reflect different mechanisms? Connectionist modeling of Dutch past-tense formation and final devoicing. *Brain and Language*, *90*, 287–298. doi:10.1016/j.bandl.2003.12.002

Peereman, R., & Content, A. (n.d.). *Quantitative analyses of orthography to phonology mapping in English and French.* Retrieved September 22, 2008, from `http://homepages.vub.ac.be/~acontent/OPMapping.html`.

Perry, C., Ziegler, J., & Coltheart, M. (2002). How predictable is spelling? Developing and testing metrics of phoneme-grapheme contingency. *The Quarterly Journal of Experimental Psychology*, *55*, 897–915. doi:10.1080/02724980143000640

Petrounias, E. V. (2002). *Neoellinikí grammatikí kai sigkritikí análisi, tómos A: Fonitikí kai eisagogí sti fonología* [Modern Greek grammar and comparative analysis, Vol A: Phonetics and introduction to phonology]. Thessaloniki: Ziti.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193. doi:10.1016/0010-0277(88)90032-7

Porpodas, C. D. (1999). Patterns of phonological and memory processing in beginning readers and spellers of Greek. *Journal of Learning Disabilities*, *32*, 40–416. doi:10.1177/002221949903200506

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.

Porpodas, C. D., Pantelis, S. N., & Hantziou, E. (1990). Phonological and lexical encoding processes in beginning readers: Effects of age and word characteristics. *Reading and Writing*, *2*, 197–208. doi:10.1007/BF00257971

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*, 143–174. doi:10.1348/000712603321661859

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423 and 623–656.

Shannon, C. E. (1950). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*, 50–64.

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. doi:10.1348/000712606X123002

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, K. A. (submitted). Predicting children's word-reading difficulty for common English words: The effect of complexity and transparency. Retrieved March 20, 2009, from `http://www.hull.ac.uk/php/edskas/`.

Stathis, C., & Carayannis, G. (1999, October). Emploutismós morfologikón lexikón me órous kai ipostíriksi kiménon entáseos óron se diadikasíes diórthosis lathón [Enriching morphological dictionaries with terms and supporting term-intensive texts in error correction procedures]. In *Praktika 2ou sinedriou "Elliniki Glossa kai Orologia"* (pp. 157–165). Athens, Greece.

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. doi:10.1037/0096-3445.124.2.107

Zevin, J. D., & Seidenberg, M. S. (2006). Simulating consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language*, *54*, 145–160. doi:10.1016/j.jml.2005.08.002

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*, 3–29. doi:10.1037/0033-2909.131.1.3

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, & Computers*, *28*, 504–515.

Ziegler, J. C., Perry, C., & Coltheart, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology*, *12*, 413–430. doi:10.1080/09541440050114570

Ziegler, J. C., Perry, C., & Coltheart, M. (2003). Speed of lexical and nonlexical processing in French: The case of the regularity effect. *Psychonomic Bulletin & Review*, *10*, 947–953.

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 inconsistency in English. *Behavior Research Methods, Instruments, & Computers*, *29*, 600–618.

Appendix
Graphophonemic conversion rules

Table A1: *Graphophonemic conversion rules, in order of rank priority. Pre: Preceding context; GRA: Grapheme; Fol: Following context; PHO: Phoneme; Opt: Optional; Reg: Regularity, i.e., proportion of rule matching opportunities, taking rank order into account, in which the rule produces the correct pronunciation. Capital latin letters refer to sets of letters (listed in Table A2) for which the rule applies.*

N°	Pre	GRA	Fol	PHO	Opt	Reg
1		αι		ɛ		1.00
2		αί		ˈɛ		1.00
3		α	ύ	ˈa		1.00
4		ά		ˈa		1.00
5		α		a		.99
6		ββ		v		1.00
7		β		v		1.00
8		γU	V	ʝ	*	.65
9		γ	H	ʝ		.92
10		γGU	V	ɣ	*	.11
11		γG	H	ɣ		1.00
12		γ	K	ŋ		1.00
13		γG		g		.99
14		γ		ɣ		1.00
15		δδ		ð		—
16		δ		ð		1.00
17	μ	U	V	ɲ	*	.13
18		ε	ύ	ˈɛ		1.00
19		η	ύ	ˈi		1.00
20		ϑ		θ		1.00
21		ϊ		i		1.00
22		ΐ		ˈi		1.00
23		κκU	V	c	*	.00
24		κU	V	c	*	.34
25		κκ	H	c		1.00
26		κ	H	c		.97
27		κκ		k		1.00
28		κ		k		1.00
29		λλU	V	ʎ	*	.43
30		λU	V	ʎ	*	.45
31		λλ		l		.99
32		λ		l		.98
33		μπ	τ	m		1.00
34		μπ		b		1.00
35		μμ		m		1.00
36		μ		m		1.00
37		ν	τZ	n		.99
38		τζ		ʤ		1.00
39		ζ		z		1.00
40		ντ		d		1.00
41		ννU	V	ɲ	*	.19
42		νU	V	ɲ	*	.48
43		νν		n		.95
44		ν	P	ŋ		1.00
45		ν		n		.99
46		ξ		ks		1.00
47		ου		u		1.00
48		ού		ˈu		1.00
49		ππ		p		1.00
50		π		p		.96
51		ρρ		ɾ		1.00
52		ρ		ɾ		1.00
53		τF		ts		1.00
54		σ	L	z		.98
55		σσ		s		1.00
56		σ		s		1.00
57		ς		s		1.00
58		ττ		t		1.00
59		τ		t		.94
60	A	Yβ		v		1.00
61	A	Yφ		f		1.00
62	A	Y	N	v		1.00
63	A	Y		f		1.00
64		ϋ		i		1.00
65		ΰ		ˈi		1.00
66		φ		f		1.00
67		χU	V	ç	*	.17
68		χ	H	ç		.99
69		χ		x		1.00
70		ψ		ps		1.00
71		ω		o		.97
72		ώ		ˈo		1.00
73	B	U	V	ç	*	.26
74	W	U	V	ʝ	*	.23
75		U		i		.95
76		T		ˈi		1.00
77		ε		ɛ		.96
78		έ		ˈɛ		1.00
79		ο		o		.99
80		ό		ˈo		1.00

Table A2: *Sets of letters defining context or graphemes for the application of graphophonemic conversion rules (Table A1).*

Symbol	Letters in set	Description
C	β, γ, δ, ζ, ϑ, κ, λ, μ, ν, ξ, π, ρ, σ, ς, τ, φ, χ, ψ, γκ, μπ, ντ, τζ	Consonants
V	ει, εί, οι, οί, υι, υί, αι, αί, ου, ού, α, ά, ε, έ, η, ή, ι, ί, ο, ό, υ, ύ, ω, ώ	Vowels
S	ϑ, κ, ξ, π, τ, σ, ς, φ, χ, ψ	Unvoiced consonants (UC)
L	β, γ, δ, ζ, λ, μ, ν, ρ, γκ, μπ, ντ, τζ	Voiced consonants (VC)
I	ει, εί, οι, οί, υι, υί, η, ή, ι, ί, υ, ύ	Vowels [i]
E	ε, έ, αι, αί	Vowels [ɛ]
U	ει, οι, υι, η, ι, υ	Unstressed vowels [i]
T	εί, οί, υί, ή, ί, ύ	Stressed vowels [i]
H	ει, εί, οι, οί, αι, αί, υι, υί, ε, έ, η, ή, ι, ί, υ, ύ	Front vowels ([i], [ɛ])
A	α, ε, η	Vowels combining with ⟨υ⟩
G	γ, κ	Consonants forming [g]
Y	υ, ύ	Vowels ⟨υ⟩
Z	σ, ς, ζ	Sibilants
B	ϑ, ξ, π, τ, σ, φ, ψ	UC with no palatal variant
W	β, δ, ζ, ρ, μπ, ντ, τζ	VC with no palatal variant
N	β, γ, δ, ζ, λ, μ, ν, ρ, γκ, μπ, ντ, τζ, α, ά, ε, έ, η, ή, ι, ί, ο, ό, υ, ύ, ω, ώ, ει, εί, οι, οί, υι, υί, αι, αί, ου, ού	Vowels and voiced consonants
P	κ, γ, χ, ξ	Velar consonants
K	χ, ξ, κτ	Velar consonants not forming [g]
F	σ, ς	Consonants [s]