Reading and Writing (2005) 18:211–229
DOI 10.1007/s11145-005-3001-9

Onset entropy matters – Letter-to-phoneme mappings in seven languages

SUSANNE R. BORGWALDT[1], FRAUKE M. HELLWIG[2] and ANNETTE M.B. DE GROOT[3]
[1]*School of Psychology, University of Sydney, Australia;* [2]*Max Planck Institute for Psycholinguistics F.C. Donders Centre for Cognitive Neuroscience, Nijmegen, The Netherlands;* [3]*Department of Psychology, University of Amsterdam, The Netherlands*

Abstract. Alphabetic orthographies vary in the (in)consistency of the relations between spelling and sound patterns. In transparent orthographies, like Italian, the pronunciation can be predicted from the spelling, in contrast to opaque orthographies such as English, where spelling–sound correspondences are often inconsistent. The pronunciation of English vowel letters is in particular very ambiguous. In this paper, we provide a cross-linguistic investigation of orthographic transparency at the word-initial letter–phoneme level, resulting in ranked metrics for the seven languages investigated: Dutch, English, French, German, Hungarian, Italian, and Portuguese, expressed as entropy values. We focus on the contributions of vowels and consonants towards the overall orthographic transparency and provide evidence that deviations from consistent word-initial 1:1 mappings between letters and phonemes influence reaction times in naming tasks. Implications for theories of visual word recognition and speech production will be discussed.

Key words: Consistency, Cross-linguistic Comparisons, Reading, Spelling

Introduction

Alphabetic orthographies, although generally based on the principle of letter–phoneme correspondences, deviate from this principle to varying degrees. Languages with a transparent (or shallow) orthography, like Italian (Maraschio, 1993), have quite predictable spelling-to-sound correspondences. In opaque orthographies, such as found in English (e.g., Carney, 1994), spelling-to-sound correspondences are often very ambiguous.

From a psycholinguistic point of view, the degree of spelling-to-sound inconsistency is known as one of the factors that affect reading performance. Words that are regularly pronounced, such as CAT, /kæt/, MILE, /maɪl/, or HINT, /hɪnt/, are read faster than words that are pronounced in a deviating way, like YACHT, /jɒt/, AISLE, /aɪl/, or PINT,

/paɪnt/. This effect – the so-called regularity effect – has been investigated in many studies researching the role of spelling-to-sound transparency in visual word recognition.

There is general consensus about the approximate classification of several languages in terms of their orthographic transparency (e.g. Seymour, Aro, & Erskine, 2003). Yet there is relatively little quantitative cross-linguistic research regarding this matter. Additionally, previous research focuses almost exclusively on monosyllables (e.g., Martensen, Maris, & Dijkstra, 2000; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995; Ziegler, Jacobs, & Stone, 1996; Ziegler, Perry, & Coltheart, 2000; Ziegler, Stone, & Jacobs, 1997). An analysis of only a relatively small and biased subset might not always result in reliable assessments of a language's overall orthographic transparency (see also Borgwaldt & De Groot, 2002).

The inconsistency of spelling-to-sound mappings can in principle be captured in various ways. Rule-approaches or analogy-based approaches can be used, the letter level, grapheme level or sublexical unit level can be analyzed, and the degree of (un)ambiguity can be expressed as a continuous or a dichotomous variable. Recent investigations into the (ir)regularity of alphabetic orthographies have focused on expressing the ambiguity in terms of entropy values. Expressing (un)ambiguities in spelling–sound mappings as entropy values will result in continuous variables, starting at zero for a totally unambiguous, predictable correspondence between spelling and sound patterns, and increasing with increasing degrees of uncertainty.

Smith and Silverberg (in revision), focused on the onsets of English words and computed entropy values for word-initial letter-to-phoneme correspondences. These calculations account for the positional restrictions of some spelling–sound combinations (e.g., <gh> is only pronounced /f/ in non-initial positions, as in LAUGH; in a word's onset it always maps to a /g/, as in GHOST, cf. Smith & Silverberg, in revision, for a detailed discussion). This context dependency might be part of the reader's knowledge, and should therefore be incorporated in analyses of languages' spelling–pronunciation mappings. A distinct advantage is that in concentrating on a word's initial part all words in that language, not only the commonly investigated monosyllabic vocabulary, can be included in an analysis of the spelling–sound correspondences. Applying Smith and Silverberg's method, Borgwaldt, Hellwig, and De Groot (2004), replicated their letter-to-phoneme analyses for English word forms, extended them by calculating phoneme-to-letter entropy values as well, and performed these bi-directional calculations for four additional languages – Dutch, German, French and Hungarian – in order

to rank these five languages on the continuum from opaque to transparent orthographies.

In this study we will calculate word-initial letter–phoneme entropies on a slightly different subset, i.e., lemmas instead of word forms, and add two new languages, Italian and Portuguese, to our comparisons. We proceed to investigate the relative contributions of languages' vowels and consonants to their overall orthographic transparency and analyze the influence of entropy values as predictors of reaction times in naming tasks.

Method and corpora

The entropy concept was introduced into the framework of Information Theory by Shannon (1948). It describes the redundancy of a communication system and is defined as follows:

For a variable x, which can take n values (y_1, y_2, ... yn) with a possibility of p, the entropy of x, $H(x)$ is the negative sum over the probability of each separate value of x multiplied by the base 2 logarithm (ld) of this probability.

If a variable takes only one value, its entropy equals 0. If a variable takes $n > 1$ values, the entropy's upper limit is ld n. That means that the more values a variable can take and the more equiprobable these values are, the higher the entropy value is.

In applying the entropy concept in order to calculate the ambiguity of word-initial letter-to-phoneme correspondences, the general idea is as follows: If a letter always corresponds to one phoneme, then its entropy will be zero, as its pronunciation is completely predictable. The more alternative pronunciations a letter has, the higher its entropy value is. In addition to the number of different pronunciations, the relative frequency of these alternative mappings contributes to the ensuing entropy value. If some of those pronunciations appear only very rarely, and if there is one truly dominant correspondence, the entropy value is lower than in the case of all pronunciations occurring with approximately the same frequency. This means that the impact of exceptional pronunciations is rather marginal. For example, in certain English words, the initial <m> is part of the grapheme <mn>, which corresponds to the phoneme /n/, like in MNEMONIC. As in most cases words starting with <m> will be pronounced with initial /m/, its entropy value, $H(m)$ remains close to zero–the value for an unambiguous correspondence.

In contrast, English words beginning with the letter <w> may be pronounced with four different phonemes–/w/, as in WHAT, /r/, as in

WRITE, /h/, as in WHOLE, and /v/, as in the German loan WELTANSCHAUUNG. Inserting the probabilities for each mapping[1] into the entropy formula, each probability is multiplied with the binary logarithm ($\log_2$) of this probability, and finally the negative sum is calculated, yielding the entropy value $H(w) = 0.482669$. As mentioned before, the higher the entropy value, the more ambiguous the variable is. If the probabilities for all the four pronunciations of the letter <w> had been the same, the entropy value would have been 2. This is considerably higher than the present $H(w)$ of 0.482669, which results from the fact that <w> has one truly dominant pronunciation, /w/, two less dominant ones, /h/, and /r/, and one marginal one, /v/.

To compute the entropy values for our seven languages, we extracted all mono- and polysyllabic lemmas from the corpora. For the three Germanic languages we used the Celex database (Baayen, Pipenbrock, & Gulikers, 1995), resulting in 41.658 words for English, 119.580 for Dutch, and 51.699 for German. Our French source was the online Lexique database, Version 1 (New, Pallier, Ferrand, & Matos, 2001), consisting of 48.337 words. Our Portuguese corpus was the Oxford Portuguese Minidictionary (Whitlam & Correia Raitt, 2002), consisting of 10.031 words. Our Italian corpus was the Moby lexicon for Italian (Ward, 1996), consisting of 60.291 lemmas, and for Hungarian we took the Multext corpus (Erjavec, 2001), consisting of 27.314 lemmas. As the Italian and Hungarian corpora did not contain phonological information, we coded the pronunciation in these two languages ourselves and let a native speaker check the generated pronunciations. As a final step, we extracted the word-initial letters/phonemes from the lemmas in our corpora with a computer program and calculated the entropy values for the different letter-phoneme correspondences.

To calculate the overall onset entropy value for a language, we summed the entropy values for all single letters weighted by their frequency of occurrence within the corpus, that is, weighted by the probability of a word beginning with this letter.

Comparing Dutch, English, French, German, Hungarian, Italian, and Portuguese according to their word-initial letter-phoneme entropy values resulted in the order depicted in Figure 1.

If an orthography were fully transparent at the word-initial letter–phoneme level – that, is, if any letter always mapped onto the same phoneme–its entropy value in Figure 1 would be zero. No orthography investigated here is that transparent. The results show clearly that when comparing word-initial letter-to-phoneme mappings, English has the most ambiguous orthography, followed by, in descending order,

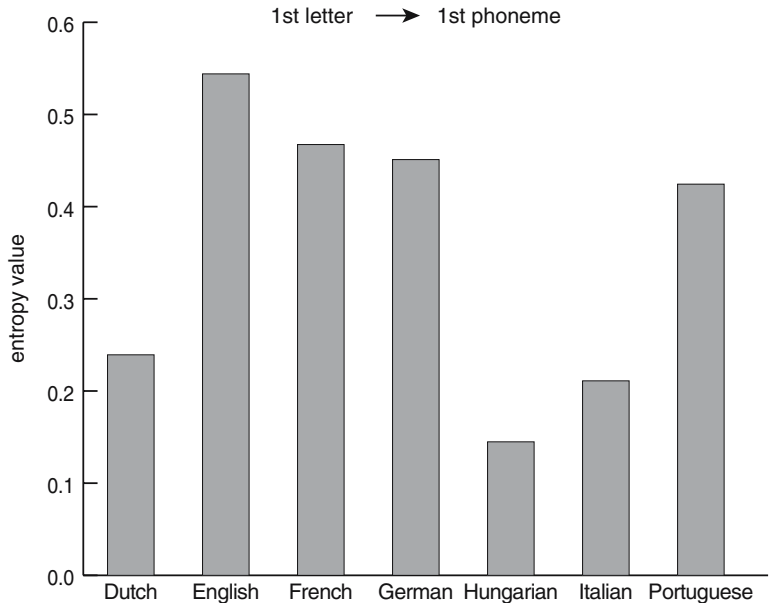

Figure 1. The relative position of the seven languages examined in terms of word-initial letter–phoneme mappings, expressed in averaged entropy values.

French, German, Portuguese, Dutch, and Italian, with Hungarian having the most predictable orthography of the seven languages examined.[2]

The entropy value for the mappings between the first letter and the first phoneme can be considered to express three distinct ambiguity components, rolled into one (Borgwaldt et al., 2004). High letter-to-phoneme entropy values could in principle reflect letter–grapheme complexities (e.g., the English letter <p> that could be part of the unambiguously pronounced grapheme <ph> as in PHILOLOGY). The degree of letter–grapheme ambiguity constitutes the first component of the overall ambiguity of first letter to first phoneme mappings. The second component that contributes to ambiguous mappings at the first letter-to-first phoneme level in isolation is the context-sensitivity of some spelling patterns. An example is the word-initial <c> in English and Dutch, whose pronunciation is quite reliably determined by the subsequent vowels (/s/ in words like CEILING and CITY, i.e., before front vowels, and /k/ in words like CAR and COPE, i.e., before back vowels). The third component is the degree of "true" letter-to-phoneme ambiguity, as for example, the German letter <v> that is pronounced like /f/ in VATER, but like /v/ in VASE.

Borgwaldt et al. (2004) also showed that when taking larger contexts, like the first two or three letters, into account and calculating the ambiguity of letter-to-phoneme mappings, the pattern of relative ambiguity of the languages as compared to one another remained about the same. At a grain size of three letters, English still remained more ambiguous than the other orthographies, whereas French, Dutch, and German approximated one another and Hungarian showed almost no ambiguities anymore.

Using the entropy method described above, the goal of this paper is to expand and complement previous cross-linguistic research on spelling–sound relations. We first investigate the distribution of entropy values in seven languages, analyzing lemmas (i.e. dictionary headwords, e.g. HUNT) instead of word forms (e.g. HUNT, HUNTS, HUNTED, HUNTING), and compare the (in) consistency of languages' vowels and consonants. Next we examine whether the ambiguities between word-initial letters and phonemes, stated as entropy values, correlate with human word-processing performance. If the generated entropy metrics impact on word processing, in other words, if the calculations based on language structures do possess descriptive adequacy for language processing, then we might expect correlations between the entropy values and reaction times in tasks that reflect word processing. If, on the other hand, higher letter-based entropy values do not correspond to longer response latencies in word-processing tasks, this might be a clue that the grain size of word processing is not letter-based, but that word processing relies on larger grain sizes such as graphemes, or on sublexical units like rimes, or even on morphemes.

Spelling-to-sound ambiguities: Vowels versus consonants

Earlier investigations into the nature of the English orthography have shown that especially the unpredictable pronunciation of vowels in isolation contributes to its high ambiguity (Brown & Besner, 1987; Treiman et al., 1995). Reasons for the larger ambiguity of vowels as compared to the ambiguity of consonants are mostly historical and lie in the imbalance between number of vowel phonemes and number of vowel letters. English, a language with over 20 vowel phonemes, monophthongs and diphthongs, faces some obvious difficulties to express them with the six vowel letters the Roman alphabet provides: <a>, <e>, <i>, <o>, <u>, and <y>.[3]

A comparison of the seven languages' consonant phoneme and vowel phoneme inventories is presented in Table 1. In this table all letters with diacritics such as accents or umlauts were counted as separate letters. Ambiguous letters, such as <y> that in German can denote vowels, like in YPSILON, as well as consonants, like in YOGHURT, were classified in this analysis according to the type of phoneme they represent the most often.

Comparing the letter and phoneme inventories in Table 1 might suggest at first sight the existence of only few spelling-to-sound ambiguities for languages like French with a relatively moderate imbalance between the letter and phoneme inventories, and more ambiguities for languages like Hungarian, where the number of phonemes is almost twice the number of letters to map onto. However, as the entropy calculations in Figure 1 show, this would be a misleading assumption. The creation of multi-letter graphemes leads to an expansion of a language's grapheme inventory and may solve existing letter–phoneme imbalances. Additionally, the mismatch between overall letter–phoneme inventories reported in Table 1 might be misleading, as in a specific position, for example, word-initially, the letter–phoneme ratio might be more balanced in some languages. On the other hand, even inventories that at first sight are quite balanced might display rather ambiguous mappings, for example due to (partly) silent letters, like <h> as in English (e.g., HOUR, /aʊə/) and French (e.g., HEURE, /œr/).

In order to further investigate the relative ambiguity for word-initial letter-to-phoneme mappings, and to provide a more accurate account of the distinct characteristics of vowels and consonants, we first counted the absolute number of word-initial letter-to-phoneme mappings. Then we calculated the average number of phonemes that the letters denote across languages in word-initial position by dividing the number of

Table 1. Letter and phoneme inventories, split up into vowels and consonants.

Language	Letters	Phonemes	Consonant letters	Consonant phonemes	Vowel letters	Vowel phonemes
Dutch	30	41	20	22	10	19
English	27	46	20	24	7	22
French	33	36	21	20	12	16
German	29	43	20	24	9	19
Hungarian	33	62	19	48	14	14
Italian	23	50	18	43	5	7
Portuguese	41	31	19	19	22	12

word-initial letter-to-phoneme mappings by the number of word-initial letters per language.

The results of these calculations are shown in Table 2.

In order to investigate the relative contributions of a language's vowels and consonants to the overall orthographic transparency expressed as entropy values we calculated separate average entropy values for the language's vowel and consonant letters. In these calculations we took the existence of ambiguous letters that denote vowels as well as consonants into account. We did this by adding all entropy values for the separate letter-to-phoneme mappings, differentiating between mappings to vowel phonemes and to consonant phonemes, and then dividing the sum by the number of letters, thus receiving an average value for the ambiguity of vowels versus consonants. To give an example, the entropy values for the letter-to-phoneme mappings of the English letter <u> that maps to a consonant, as in UNIFORM, or to a vowel, as in UNDER, were split up and counted separately. In the above mentioned example the overall entropy value of <u>, H(u) equals 0.695379. From this value 0.340729 contributed to the vowel entropy, and 0.354650 to the consonant entropy. This is illustrated in Table 3.

The results of these calculations are presented in Figure 2

Here we can see striking differences between vowel entropy values and consonant entropy values, reflecting the obvious imbalance between vowel phonemes and vowel letters[4]. For vowels, English is the language with the most ambiguous orthography-to-phonology mappings, followed by, in decreasing order, German, Dutch, French, Portuguese, Italian, and Hungarian. Of the seven languages, Hungarian is the only one that shows no ambiguities for vowels, presumably resulting from the fact that Hungarian does not contain diphthongs in the vowel

Table 2. Word-initial letter-to-phoneme mappings: letter-to-phoneme ratio.

Language	# of word-initial letter-to-phoneme mappings	Phoneme average per letter
Dutch	80	2.666666
English	105	3.888888
French	94	2.848484
German	77	2.655172
Italian	50	2.173913
Hungarian	30	0.909090
Portuguese	74	2.176470

Table 3. Examle of vowel and consonant entropy values for an ambiguous initial letter.

1st letter	1st phoneme	Examples	Probability	Entropy value for vowels	Entropy value for consonants
u	3:	urban	0.015081	0.091258	
u	ʊə	urdu	0.002320	0.020305	
u	ə	until	0.004640	0.035970	
u	ʊ	umlaut	0.001160	0.011313	
u	Λ	ugly	0.864269	0.181883	
			total: 0.887471	total: 0.340729	
u	J	uniform	0.112529		0.354650
			total: 0.112529		total: 0.354650

phoneme inventory. Diphthongs are often denoted with digraphs, and would therefore cause ambiguity at the first letter to first phoneme level. In two languages, Hungarian and Italian, the numbers of vowel letters and vowel phonemes are quite balanced, as has been shown in Table 1. In Hungarian all vowel phonemes are differentiated by diacritics, e.g., <u>, <ú>, <ü>, and <ü>. In this way every vowel

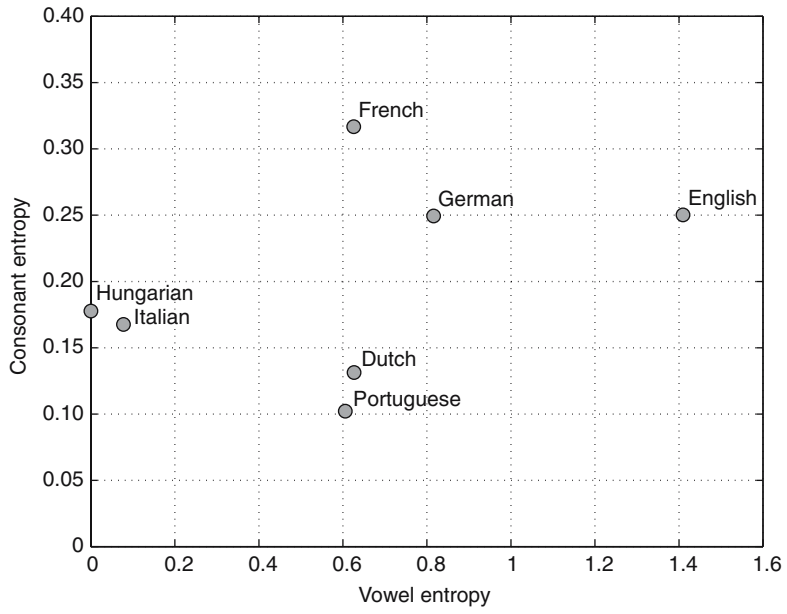

Figure 2. Initial letter-to-phoneme average entropy values for vowels and consonants.

phoneme is denoted by a distinct vowel letter. In Italian only two vowel letters, <e> and <o> can correspond to two different phonemes.

For consonants, the pattern changes: Here, French is the language with the most ambiguous orthography-to-phonology mappings, followed by, in decreasing order, English, German, Hungarian, Italian, Dutch, and Portuguese. Two languages, Italian and Hungarian, showed greater entropy values for consonants than for vowels. Examples of ambiguous consonant letter to consonant phoneme mappings in Hungarian followed from the existence of digraphs and trigraphs in Hungarian. For example, word-initial <s> can be part of the digraph <sz>, /s/, and word-initial <d> can be part of the trigraph <dzs>, /d_3/. In Italian, like in Hungarian, ambiguous consonant letter to consonant phoneme mappings resulted from the existence of digraphs and trigraphs, or, like in Dutch or English, from the existence of context-dependent phonemes. For example, analogous to the context dependency of the letter <c> in English, also in Italian the pronunciation of the letters <c> and <g> depends on the following phonemes.

The results presented above show the differently distributed orthography-to-phonology ambiguities for vowels and consonants in the seven languages investigated. In addition to providing more differentiated descriptions of languages' orthographic transparency, the obtained results might be useful for further validating reading models that propose separate phonological activation processes for vowels versus consonants. In Berent and Perfetti's (1995) two-cycles model of phonology, based on autosegmental theories of phonology (e.g., Goldsmith, 1990), an English word's pronunciation is derived using two distinct processes. First, consonant phonemes are assembled by an automatic computational mechanism, and, later, vowel phonemes are added in a separate slower process. Empirical support for this hypothesis was presented in backward masking, naming, and fast priming paradigms, where consonant-preserving conditions produced higher accuracies than vowel-preserving conditions at very brief presentations (Berent & Perfetti, 1995; Lee, Rayner, & Pollatsek, 2001, 2002). However, Colombo, Zorzi, Cubelli, and Brivio (2003) found the opposite pattern of results for Italian, that is, they obtained a processing priority for vowels in Italian. They argued that these differences in consonant–vowel processing across languages might suggest that this process is not a structural hypothesis (i.e., that consonants and vowels are represented phonologically in different ways) but just reflects language-specific characteristics (i.e., different C-V properties such as spelling-to-sound ambiguity), and is, as such, only a statistical hypothesis. In the statistical analyses of spelling-to-sound ambiguities reported above the finding emerged that

especially in English many letters were ambiguous in terms of consonant–vowel status, that is, they could denote consonants as well as vowels. Therefore it seems unlikely that inherently ambiguous items could be encoded as distinct linguistic entities, like the Berent and Perfetti model predicts. We suggest, in line with Lee et al. (2001), that the results obtained by Berent and Perfetti can best be explained in terms of a statistical hypothesis. Our results presented above can then be used to predict the behavior of the other languages in terms of their consonant–vowel encoding.

To summarize, the distinct vowel–consonant measures and the onset entropy calculations seem to provide quite reliable estimations of the languages' orthographic transparency, resulting in rankings that are in line with earlier descriptions of the languages' spelling and pronunciation patterns (cf. Carney, 1994; Eisenberg, 1998; Maraschio, 1993; Nunn, 1998; Peereman & Content, 1999; Siptár & Törkenczy, 2000). As further analyses revealed, the seven languages analyzed displayed different characteristics in terms of vowel versus consonant ambiguity. These characteristics can explain language-specific behavior of phonological encoding during the reading process.

In order to test whether the degree of (un)ambiguity of the letter-to-phoneme correspondences affects actual language processing, we proceeded to investigate the influence of the onset entropy values on naming latencies. If, instead of relying on letter–phoneme mappings, language users rely exclusively on other, larger, functional reading units like graphemes, sublexical units, or morphemes, they might not be impaired by ambiguities at the letter level.

The role of onset entropy in word naming

Spelling-to-sound ambiguities are known to influence reaction times in reading tasks (e.g., Glushko, 1979, Treiman et al., 1995). The choice of spelling-to-sound variables with the goal to use them as predictors for human performance should aim at capturing ambiguities that influence human performance. As a consequence of this principle, an analysis of the sub-parts of words should concentrate on those sub-parts whose spelling–sound irregularities influence human performance the most.

While for this reason the majority of researchers (e.g., Glushko, 1979; Ziegler et al., 1996, 1997) have focused on rime analyses (of monosyllabic words), another candidate for a detailed investigation of spelling-to-sound mappings is the beginning of a word. Treiman et al. (1995) showed that ambiguities in (consonantal) onsets accounted for

more of the variance in reaction times than ambiguities in (consonantal) codas.

That the word's onset plays a crucial role in lexical access tasks has been demonstrated in research on spoken word recognition. For instance, studies testing the cohort model of spoken word recognition (Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Zwitserlood, 1989) predict sequential processing and stress the importance of word onsets. In addition, there is considerable evidence that also in written word recognition tasks word onsets are salient and disproportionately important identification units (cf. Cutler, 1982; Rayner & Pollatsek, 1989; Smith & Silverberg, in revision). An inherent advantage of looking at word-initial spelling–sound relations lies in the relative stability of pronunciation of the word's onsets. A study by Greenberg, Carvey, Hitchcock, and Chang (2002), comparing pronunciation in spontaneous speech with canonical pronunciation, revealed that syllable nuclei or codas were much more likely to be deleted than syllable onsets, that, in contrast, were almost always realized.

At least some models of reading, for instance, dual-route models (Coltheart, 1978), assume serial processing, from left to right, or contain serial parsing components. Such processing consequently predicts a larger impact on reaction times by ambiguities at the beginning of a word than by ambiguities at the end of a word. This prediction is supported by human performance in word naming tasks, resulting in a positional regularity effect, that is, the effect of irregularity decreases the later the position of the irregularity in the word to be read. Ambiguities at the beginning of a word cause a larger delay than ambiguities at the end of a word, because in the latter case the pronunciation output from the lexical route, that operates in parallel to the nonlexical route, may already have been delivered by the system before the ambiguity is actually encountered (Coltheart & Rastle, 1994).

The positional regularity effect has been observed in studies by Coltheart and Rastle (1994) and has been replicated by Cortese (1998), and by Rastle and Coltheart (1999). However, Zorzi (2000) showed that connectionist models could also account for a positional regularity effect without assuming serial components.

In order to test the impact of onset entropy values cross-linguistically, we correlated the onset letter–phoneme entropy values, that is, the entropy values for the mappings between the first letter and first phoneme, with reaction times of large scale word naming studies carried out in three of the seven languages. The languages were Italian, with a shallow orthography (Barca, Burani, & Arduino, 2002), Dutch, with an

intermediate orthography, and English, with an opaque orthography (De Groot et al., 2002). The Italian study reports mean response latencies of 626 polysyllabic Italian nouns, collected on 30 participants. The Dutch and English study reports mean response latencies of 440 mono- and polysyllabic Dutch and English nouns, collected on 40 participants for each language.

Each word was assigned the entropy variable for its first letter, expressing the general degree of the letter's ambiguity, irrespective of the probability of the specific mapping between the first letter and the first phoneme. For example, all words starting with the letter <a> received the same entropy value, regardless of (the probability of) their pronunciation. This is illustrated in Table 4.

As the naming latencies had been recorded with voice keys, it is likely that they have responded to specific articulatory features of the word's beginnings such as its sound intensity (Kessler, Treiman, & Mullennix, 2002; Rastle & Davis, 2002). In our study we wanted to attenuate such voice key effects and other possibly confounding effects from the effects that we were primarily interested in, that is, the effects of letter-to-phoneme ambiguity, expressed as entropy value. Of the variables studied by De Groot et al. (2002), two correlated significantly with the present entropy variables, namely, consonant cluster structure and sound intensity. Consonant cluster structure concerned the number of word-initial consonant phonemes, and the sound intensity variable coded the initial phonemes according to their average sound intensity in decibels (see De Groot et al., 2002, for further details of these two variables, which are referred to as ONS2, and INT1 in that study). We removed the effects of these two confounding variables by computing partial correlations between the mean reaction times and the entropy values of the initial letter-to-phoneme mappings, partialling out consonant cluster structure and sound intensity.[5]

Table 4. Entropy values assigned to the words – independent of the probability of the specific mapping/pronunciation.

word	1st phoneme	Probability of this mapping	Entropy value
abuse	/ə/	0.387927	0.482669
action	/æ/	0.042634	0.482669
advantage	/ə/	0.387927	0.482669
age	\eɪ\	0.042634	0.482669

English sample from the stimuli used in De Groot et al. (2002).

For all three languages, of varying orthographic transparency, the partial correlations were significant: The higher the onset entropy, the longer the reaction times. This is shown in Table 5.

It is a well known fact that spelling-to-sound ambiguities do affect reaction times in reading tasks. However, the majority of relevant earlier studies could only demonstrate the impact of spelling-to-sound ambiguities for languages with a rather opaque orthography, for example English (Treiman et al., 1995) and French (Lange & Content, 1999). Therefore, the most noteworthy aspect of the present finding is the fact that the effects of onset letter-to-phoneme ambiguity can also be shown for languages with a very shallow orthography, as we demonstrated for the Italian data set.

The present effect of word-initial entropy on naming times could also provide some insight into the time course of word naming, an issue investigated in the area of speech production. According to the initial-phoneme criterion (e.g. Kawamoto, Kello, Jones, & Bame, 1998) the articulation of a (monosyllabic) word begins as soon as the phonology for the initial phoneme has been generated. According to the whole word criterion (e.g. Rastle, Harrington, Coltheart, & Palethorpe, 2000), a complete phonological representation of the whole word is required before the articulation is generated. Any effect of word-initial spelling–sound inconsistency on reaction times in naming tasks supports the initial-phoneme criterion, i.e. that the phonology does not have to be generated completely before articulation begins.

Conclusions

We investigated ambiguities in word-initial letter-to-phoneme mappings in seven languages in detail concentrating on differences in vowel versus consonant entropy values and demonstrated significant correlations between the descriptive statistics and naming latencies in three of these languages.

Table 5. Partial correlations between letter-to-phoneme entropy values and naming latencies in three languages of varying orthographic transparency.

Language	Partial correlation coefficient (letter-phoneme entropy–RT's)	*P*-value (two-tailed)
Italian	0.1806	< 0.001
Dutch	0.1706	< 0.001
English	0.2424	< 0.001

All orthographies examined deviate to various degrees from the "ideal" one-to-one mapping between letters and phonemes. In line with earlier comparisons between a subset of these languages (Martensen et al., 2000; Van den Bosch, Content, Daelemans, & de Gelder, 1995; Ziegler et al., 1997) we have found that in terms of overall spelling-to-sound relations examined at the word-initial letter–phoneme level, English has the most ambiguous orthography, followed by (in decreasing order), French, German, Portuguese, Dutch, Italian, and Hungarian. The pattern changes slightly when consonant and vowel letters are analyzed separately. For vowels, English remains the language with the most ambiguous letter-to-sound relations, followed by (in decreasing order) German, Dutch, French, Portuguese, Italian, and Hungarian. Hungarian shows completely unambiguous vowel letter-to-vowel phoneme mappings. For consonants, French shows the highest letter-to-phoneme ambiguity, followed by (in decreasing order) English, German, Hungarian, Italian, Dutch, and Portuguese. None of the studied orthographies displayed completely unambiguous mappings between consonant letters and consonant phonemes.

The onset entropy calculations as described above not only appear to be a method to become informed on languages' overall orthographic transparency, they also provide a way to classify single words according to the degree of spelling-to-sound ambiguity of their word-initial letters, a variable that, as demonstrated here, correlates significantly with naming latencies.

A clear effect of deviations from a one-to-one mapping between word-initial letters and phonemes as expressed in entropy values on naming was found. Even in very transparent orthographies (like Italian) and also in languages of intermediate (Dutch) and opaque (English) orthographic transparency, these ambiguities influence reaction times in naming tasks. This suggests that the ambiguity of letter–phoneme mappings cannot be ignored in favor of an exclusive focus on larger grain sizes like sublexical units or graphemes, where most of the current research is focusing on. Additional support for this view comes from the "whammy effect" demonstrated by Rastle and Coltheart (1998). They studied nonword naming, assuming a dual route model, and concluded that the functional reading unit of the indirect route is the letter, as it took their participants longer to name nonwords with phonemes mapping to digraphs or trigraphs than nonwords with a one-to-one mapping of letters to phonemes. That is, monosyllabic nonwords consisting of five letters and five phonemes, such as TRUSP, caused shorter naming times than monosyllabic nonwords consisting of five letters but three phonemes, such as FOOCE. These findings provide a seri-

ous challenge for reading theories that postulate only sublexical units as functional reading units, and for models of reading that do not allow for any graded effects of ambiguities.

However, Lange and Content (2000), showed that the ''whammy effect'' could be due to a confound with grapheme frequency, and other researchers continue to provide evidence for graphemes instead of letters as perceptual reading units, e.g. Rey, Ziegler and Jacobs, (2000). Although the significant correlations between word-initial letter–phoneme entropy values and reaction times that we found are consistent with the assumption that letters are the functional reading units, this might partly follow from a correlation between letter–phoneme ambiguity and grapheme–phoneme ambiguities.

An interesting future line of research we would like to pursue is to see whether and how entropy values for larger grain sizes, and the reverse entropy values, that is, in sound-to-spelling direction, relate to the statistical and behavioral data. Analogously to the correlations between letter-to-phoneme entropies and reaction times, reported here, we can then explore the influence of grapheme-to-phoneme, phoneme-to-letter, and phoneme-to-grapheme entropies on reaction times, in order to investigate possible feedback effects.

Notes

1. The entropy values reported in this article are based on type frequencies. However, the general pattern of results stayed the same when we calculated entropy values based on token frequencies.
2. These results are roughly comparable to Borgwaldt et al. (2004), who analyzed the onset entropy of Dutch, English, German, French and Hungarian word forms.
3. The status of hybrid, ambiguous letters that can denote vowels as well as consonants will be discussed in more detail in the next section.
4. Note, that in the computations above the relative frequency of word-initial letters was not taken into account, that is, we averaged across letters without taking into account that certain letters might occur in word-initial position only very rarely, whereas other letters might occur very often in that position.
5. However, with this method we could not correct for all phonetic biases, as Kessler et al. (2002) showed that for example even the acoustic properties of the second phonemes influence reaction times.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database*. (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, and Computers, 34*(3), 424–434.

Berent, I., & Perfetti, C. A. (1995). A rose is a REEZ: The two cycles model of phonology assembly in reading English. *Psychological Review, 102*, 146–184.

Borgwaldt, S. R., & De Groot, A. M. B. (2002). Beyond the rime–measuring consistencies for monosyllabic and polysyllabic words. In M. Neef, A. Neijt, & R. Sproat (Eds.), *The relation of writing to spoken language*. (pp. 49–69). Tübingen: Niemeyer.

Borgwaldt, S. R., Hellwig, F., & De Groot, A. M. B. (2004). Word-initial entropy in five languages - letter to sound and sound to letter. *Written Language and Literacy, 7* (2), 165–184.

Brown, P., & Besner, D. (1987). The assembly of phonology in oral reading: a new model. In M. Coltheart (Ed.), *Attention and Performance XII: The Psychology of Reading*. (pp. 471–489) Hillsdale, NJ: Erlbaum.

Carney, E. (1994). *A survey of English spelling*. London: Routledge.

Colombo, L., Zorzi, M., Cubelli, R., & Brivio, C. (2003). The status of consonants and vowels in phonological assembly: Testing the two-cycles model with Italian. *The European Journal of Cognitive Psychology, 15*(3), 405–433.

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

Coltheart, M., & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1197–1211.

Cortese, M. J. (1998). Revisiting serial position effects in reading. *Journal of Memory and Language, 39*(4), 652–665.

Cutler, A. (1982). The reliability of speech error data. In A. Cutler (Ed.). *Slips of the tongue and language production*. (pp. 7–28). Amsterdam: Mouton.

De Groot, A. M. B., Borgwaldt, S., Bos, M., & van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language, 47*(1), 91–124.

Eisenberg, P. (1988). Die Grapheme des Deutschen und ihre Beziehung zu den Phonemen. *Germanistische Linguistik, 93/94*, 139–154.

Erjavec, T. (2001). Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984, 6th Natural Language Processing Pacific Rim Symposium, Tokyo, NLPRS'01 (pp. 487–492)[Online] http://nl.ijs.si/ME/.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 674–691.

Goldsmith, J. (1990). *Autosegmental and metrical phonology*. Cambridge: Blackwell.

Greenberg, S., Carvey, H. M., Hitchcock, L., & Chang, S. (2002). Beyond the phoneme - A juncture-accent model of spoken language. *Proceedings of the Second International Conference on Human Language Technology Research*. (pp. 36–44) [Online] http://www.icsi.berkeley.edu/~steveng/PDF/Beyond_the_Phoneme.pdf

Kawamoto, A. H., Kello, C. T., Jones, R. J., & Bame, K. (1998). Initial phoneme versus whole word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 862–885.

Kessler, B., & Treiman, R. (2001). Relations between sounds and letters in English monosyllables. *Journal of Memory and Language, 44*(4), 592–617.

Kessler, B. Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language, 47*(1), 145–171.

Lange, M., & Content, A. (2000). Grapheme complexity and length effects in visual word recognition. Paper presented at the 41st Meeting of the Psychonomic Society, New Orleans, Louisana, USA.

Lee, H., Rayner, K., & Pollatsek, A. (2001). The relative contribution of consonants and vowels to word recognition during reading. *Journal of Memory and Language, 44*, 189–205.

Lee, H., Rayner, K., & Pollatsek, A. (2002) The processing of consonants and vowels in reading: Evidence from the fast reading paradigm. *Psychonomic Bulletin and Review, 9*(4), 766–772.

Maraschio, N. (1993). Grafia e ortografia: evoluzione e codifiazione. In L. Seriami & P. Trifone (Eds.), *Storia della lingua italiana*. (vol. I, pp. 139–227) Turin: Guilio Einaudi Editore.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*, 71–102.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 576–585.

Martensen, H., Maris, E., & Dijkstra, T. (2000). When does inconsistency hurt? On the relation between consistency effects and reliability. *Memory & Cognition, 28*, 648–656.

New B., Pallier C., Ferrand L., & Matos R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Psychologique, 101*, 447–462. [Online] http://www.lexique.org

Nunn, A. M. (1998). *Dutch orthography. A systematic investigation of the spelling of Dutch words*. The Hague: Holland Academic Graphics.

Peereman, R., & Content, A. (1999). LEXOP: A lexical database providing orthography-phonology statistics for French monosyllabic words. *Behavior Research Methods Instruments & Computers, 31*, 376–379.

Rastle, K., & Coltheart, M. (1998). Whammy and double whammy: Length effects in nonword naming. *Psychonomic Bulletin and Review, 5*, 277–282.

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

Rastle, K., & Davis, M. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 307–314.

Rastle, K., Harrington, J., Coltheart, M., & Palethorpe, S. (2000). Reading aloud begins when the computation of phonology is complete. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1178–1191.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, NJ: Lawrence Erlbaum.

Rey, A., Ziegler, J. C., & Jacobs, A. M. (2000). Graphemes are perceptual reading units. *Cognition, 75*(1), B1–B12.

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

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

Siptár, P., & Törkenczy, M. (2000). *The phonology of Hungarian*. Oxford: Oxford University Press.

Smith, J., & Silverberg, N. (in revision). *Frequency and correspondences of initial letter/phoneme combinations for English words*. [Online] http://www.public.asu.edu/~jfsmit1/pages/letrpho.html

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.

Van den Bosch, A., Content A., Daelemans, W., & de Gelder, B. (1995). Measuring the complexity of writing systems. *Journal of Quantitative Linguistics, 1*(3), 178–188.

Ward, G. (1996). The Moby lexicon project. [Online] ftp://ftp.dcs.shef.ac.uk/share/ilash/Moby/mlang.tar.Z

Whitlam, J., & Correia Raitt, L. (2002). *Oxford Portuguese minidictionary*. Oxford: Oxford University Press.

Ziegler, J. C., Jacobs, A. M., & Stone, G. O. (1996). Statistical analysis of the bidirectional inconsistency of spelling and sound in French. *Behavior Research Methods, Instruments, and Computers, 28*(4), 504–515.

Ziegler, J. C., Stone, G. O., & Jacobs, A. M. (1997). What's the pronunciation for –ough and the spelling for /u/? A database for computing feedforward and feedback inconsistency in English. *Behavior Research Methods, Instruments, and Computers, 29*(4), 600–618.

Ziegler, J. C., Perry, C., & Coltheart, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *The European Journal of Cognitive Psychology, 12*(3), 413–430.

Zorzi, M. (2000). Serial processing in reading aloud: No challenge for a parallel model. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 847–856.

Address for correspondence: Dr. S. R. Borgwaldt, University of Alberta, Department of Linguistics, 4-32 Assiniboia Hall, Edmonton, Alberta, Canada T6G 2E5,
E-mail: suske7@yahoo.com