Word-Initial Entropy in Five Languages: Letter to Sound, and Sound to Letter

Article *in* Written Language & Literacy · January 2004

DOI: 10.1075/wll.7.2.03bor

3 authors, including:

Frauke Hellwig
Radboud University
13 PUBLICATIONS **654** CITATIONS
SEE PROFILE

Annette M. B. De Groot
University of Amsterdam
75 PUBLICATIONS **5,008** CITATIONS
SEE PROFILE

Word-initial entropy in five languages

Letter to sound, and sound to letter

Susanne R. Borgwaldt, Frauke M. Hellwig and
Annette M. B. de Groot
University of Sydney / Max-Planck-Institut for psychologuistics
and F. C. Donders Centre for Cognitive Neuroimaging Nijmegen /
University of Amsterdam

Alphabetic orthographies show more or less ambiguous relations between spelling and sound patterns. In transparent orthographies, like Italian, the pronunciation can be predicted from the spelling and vice versa. Opaque orthographies, like English, often display unpredictable spelling–sound correspondences. In this paper we present a computational analysis of word-initial bi-directional spelling–sound correspondences for Dutch, English, French, German, and Hungarian, stated in entropy values for various grain sizes. This allows us to position the five languages on the continuum from opaque to transparent orthographies, both in spelling-to-sound and sound-to-spelling directions. The analysis is based on metrics derived from information theory, and therefore independent of any specific theory of visual word recognition as well as of any specific theoretical approach of orthography.

1. Introduction

Alphabetic orthographies of languages, though all based on the principle of letter–phoneme correspondences, are more or less "regular", that is, they display more or less ambiguous relations between spelling and sound patterns. In languages with a transparent (or shallow) orthography, like Finnish (Karlsson, 1983) or Italian (Maraschio, 1993), the pronunciation can be predicted from the spelling and vice versa. In contrast, in languages with an opaque (or deep) orthography, like Danish (Gottlieb, 2001) or English (Carney, 1994; Venezky, 1970; Wijk, 1966), spelling-to-pronunciation mappings are often quite unpredictable and ambiguous. In other words, alphabetic orthographies display not always *one-to-one* correspondences between letters and phonemes, but

Written Language & Literacy 7:2 (2004), **165–184**.
ISSN 1387–6732 / E-ISSN 1570–6001

to varying degrees also *one-to-many*, *many-to-one*, and *many-to-many* correspondences.

The (un)predictability of languages' writing systems has been researched from different perspectives. Apart from computational linguists, theoretical linguists, and education specialists, psycholinguists and cognitive psychologists have focused on investigating the role of spelling-to-sound transparency in visual word recognition. In some psycholinguistic frameworks, for example in "dual-route models" (Coltheart, 1978), different processing mechanisms are assumed for regularly versus irregularly spelled words. In this theory, words can be identified in two ways, via an indirect route, for infrequent and/or regularly spelled words, by assembling the pronunciation of a word using grapheme–phoneme-correspondence rules, or via a direct route, for frequent and/or irregularly spelled words, by accessing the words' meaning without the use of phonology. An extension of this model, the *orthographic depth hypothesis* (Frost, Katz, & Bentin, 1987), proposes different processing mechanisms for deep vs. shallow orthographies. This hypothesis assumes that reading in a deep orthography, like Hebrew, relies on the direct route, whereas lexical processing in a shallow orthography, like Serbo-Croatian, involves primarily the indirect route.

There is general consensus about the approximate classification of a language in terms of orthographic transparency. Yet there is a lack of quantitative cross-linguistic research regarding this matter. In addition, psycholinguistic investigations into a language's orthography mainly concern the relation from spelling to sound patterns (*feedforward*); considerably less research has been done in the opposite (*feedback*) direction. Exceptions are, for example, the studies by Ziegler, Jacobs, and Stone (1996) and Ziegler, Stone, and Jacobs (1997), who analyzed bi-directional correspondences of both English and French rimes and concluded that reading performance is not only influenced by spelling-to-sound irregularities and inconsistencies but also by irregularities and inconsistencies in the other direction.

Another limitation lies in the psycholinguistic focus on monosyllabic and monomorphemic words (e.g., Martensen, Maris, & Dijkstra, 2000; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995; Ziegler et al., 1996, 1997; Ziegler, Perry, & Coltheart, 2000). The percentage of monosyllabic words varies across languages, and their characteristics may deviate from those of polysyllabic words in the language in terms of, for example, etymology, and, consequently, orthography. At least for some languages an analysis of only this part of the vocabulary may not generate reliable results for estimations of the

overall orthographic transparency, as we cannot generalize from samples that small and biased (see also Borgwaldt & De Groot, 2002).

This shows, in our view, the need for further cross-linguistic investigations into languages' orthographic transparency, better tailored to psycholinguistic purposes. What is needed is a detailed analysis of languages' spelling-to-sound relations, informed by and informing theories of reading, and providing valid estimates of a language's overall orthographic transparency as well as giving a detailed account of (un)ambiguities for different grain sizes.

For the sake of simplicity, we will not make a distinction between a *regular spelling* and a *regular pronunciation*, but conflate these two notions. Because a strictly synchronic point of view is adopted in the analyses to be reported below, we cannot determine whether, for example, *aisle*, /aɪl/, is *irregularly pronounced* or *irregularly spelled*. As a consequence, we will treat both the (un)ambiguity of spelling-to-sound mappings, and the (un)ambiguity of sound-to-spelling mappings as aspects of a language's orthography, and not its phonology.

The ambiguity of spelling-to-sound mappings can in principle be captured in various ways. One may simply count the number of different pronunciations that correspond to a certain spelling pattern, or compute the probabilities for the different mappings (cf. Dewey, 1971, Berndt, Reggia, & Mitchum, 1987, for descriptions of English). In contrast to these graded measures we can also interpret the ambiguity of mappings as a categorical dichotomous variable, either *ambiguous* or *unambiguous*. This is a technique that is, for instance, used within the consistency paradigm (cf. Ziegler et al., 1996, 1997). Alternatively, one may measure the irregularities with a rule-approach by defining spelling–sound correspondence rules and calculate the number of items that do not conform to the rules. Examples are the study by Ziegler et al. (2000) who investigated the regularity of German monosyllabic words in this way, or the cross-linguistic comparison of grapheme–phoneme rules defining the pronunciation of French, English and German monosyllabic words by Coltheart (2003). Another possibility is to look for the average grain size that is required as disambiguating context. This was, for instance, done by Van den Bosch, Content, Daelemans, and de Gelder (1995), who compared Dutch, English and French in terms of grapheme–phoneme mappings and letter–grapheme complexity.

Recent investigations into the (ir)regularity of alphabetic orthographies have focused on expressing the ambiguity in terms of "entropy" values. The entropy concept was introduced into information theory by Shannon (1948) and is described in more detail in our *Method* section. Expressing (un)ambiguities in spelling-to-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.

Treiman et al. (1995) calculated spelling-to-sound entropy values of English monosyllabic CVC words. They focused on the correspondences between the sub-lexical units *onset* (the initial consonant(s) of a monosyllabic word), *nucleus* (the vowel or diphthong), *coda* (the final consonant(s)), and their combinations *body* (i.e., onset and nucleus) and *rime* (i.e., nucleus and coda). Martensen et al. (2000) replicated this analysis for Dutch, showing that Dutch is indeed more regularly pronounced than is English. Lange and Content (1999), in contrast, concentrated on the grapheme level and computed grapheme-to-phoneme entropy values for French.

Smith and Silverberg (in revision), also using an entropy-based method, were the first psycholinguists to combine the positional restrictions accounted for by a sub-lexical unit approach and the focus on the letter level. They did so by first differentiating between word-initial and word-non-initial letter-to-phoneme mappings and then computing word-initial letter-to-phoneme entropy values for English. These calculations take into account the fact that some spelling–sound combinations occur only in a certain context (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). They also acknowledge that such positional restrictions might be part of a reader's knowledge, and should therefore be reflected in an analysis of a language's spelling-pronunciation mappings.

A possible disadvantage of their approach for cross-linguistic comparisons could be that the results based on only a part of a word, in this case its initial letters/phonemes, might not generalize to the remaining parts of the words, that is, to a global word level. But this assumption would require that in some languages word-initial letter–phoneme relations deviate more from non word-initial letter–phoneme relations, than they do in other languages. To our knowledge, there is no indication that the languages analyzed in this study differ significantly in the distribution of spelling–sound ambiguities within words.

However, a distinct advantage of the method is that in concentrating on word-initial sound–spelling correspondences *all* words in a language, not only its monosyllabic vocabulary, can be included in an analysis of the language's orthographic transparency.

The goal of this study is to provide a statistical description of the orthographic transparency of five languages as derived from analyses of these languages' complete vocabulary, that is, without any restrictions concerning, for instance, the words' syllable structure, phonology, morphology, or etymology. For these reasons we have applied Smith and Silverberg's (in revision) method, replicating their letter-to-phoneme analyses for English and extending them by calculating phoneme-to-letter entropy values as well. The latter is done in order to gain also an insight into a language's sound-to-spelling patterns. By performing these calculations for five languages we can determine their relative orthographic transparency with respect to one another.

2. Characteristics of the languages investigated

Five languages belonging to three different language groups and all using versions of the Latin alphabet are investigated with respect to their bi-directional sound-spelling relations, taking a strictly synchronic point of view.

Three of these languages, Dutch, English, and German, belong to the Germanic language group; together with Afrikaans and Frisian they form its West-Germanic branch. All three languages can be categorized as fusional languages (i.e., words can consist of more than one morpheme and morphemes are not easily segmentable), though English especially, and Dutch to a lesser degree, also display isolating tendencies (i.e., single words correspond to single morphemes and vice versa). Their respective phoneme inventories in word-initial position consist of 39 phonemes for Dutch, 46 phonemes for English and German (Baayen, Piepenbrock, & Gulikers, 1995). For all three languages, word stress is not fully predictable, though it mainly falls on the first syllable of the word stem (König & van der Auwera, 2003).

French, a Romance language, belongs also to the fusional languages. Its phoneme inventory contains 35 word-initial phonemes. Word stress in French is fully predictable; it always falls on the final syllable (Harris & Vincent, 1990; Posner, 1996).

The remaining language, Hungarian, is an example of a Finno-Ugric language. It is an agglutinative language (i.e., words can consist of more than one morpheme and morphemes are easily segmented), displaying vowel harmony. Its phoneme inventory consists of 39 word-initial phonemes. Hungarian's word stress is always on the first syllable (Abondolo, 1998; Siptár & Törkenczy, 2000; Vago, 1980).

There is general consensus about the approximate classification of these five languages in terms of orthographic transparency. Hungarian (Siptár & Törkenczy, 2000) is a prime example of an orthographically regular language, English is a clear case of a language with an irregular orthography (Carney, 1994, 1997; Venezky 1970, 1999), and the remaining three languages, French (Véronis, 1986), Dutch (Nunn, 1998), and German (Eisenberg, 1998) fall somewhere in between. French sound–spelling relations are considered to be quite asymmetrical: French spelling-to-sound relations are reasonably predictable, but French sound-to-spelling mappings are considered to be fairly ambiguous (Ziegler et al., 1996).

This rough classification of the languages' spelling–sound relations is easily supported by a non-scientific observation: A brief look at comparable bilingual dictionaries of the five languages shows interesting differences, that presumably originate from the common sense principle in publishing that redundant information should not be incorporated into reference books. In dictionaries for Hungarian almost never any phonetic transcriptions are included. The words' pronunciation can easily be derived from a small table of grapheme-to-phoneme rules, commonly found at the beginning of the book. On the other hand, there are almost no dictionaries for English without phonetic transcriptions. French, Dutch, and German seem to lie somewhere in between: There are dictionaries with pronunciation information for every word, and there are the ones that give only the pronunciation for exception words like foreign loans. Generally, however, in German and Dutch dictionaries word stress is indicated, but not in French dictionaries, since, as already mentioned above, it is totally predictable in French, but not in German or Dutch. Additionally, vowel length is often marked in German and not in Dutch dictionaries, presumably because this source of ambiguities exists for German pronunciation, but not for Dutch. The existence of these ambiguities in German suggests that German pronunciation is somewhat more irregular than Dutch pronunciation (but see Seymour, Aro, and Erskine, 2003, for opposing views).

3. Method

The entropy concept was first used in 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 ($\mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_n$) with a probability of **p**, the entropy of **x**, **H(x)** is:

$$H(x) = -\sum_{i=1}^{n} p_i (x = y_i) \cdot \mathrm{ld}\,(p_i(x = y_i))$$

H is thus the negative sum over the probability of each separate value of x multiplied by the base 2 logarithm (ld) of this probability. This is because the binary logarithm of a proportion is a negative number, e.g., ld(1/4) = −2.

If a variable takes only one value, its entropy equals 0. If a variable takes $n > 1$ values, the entropy's upper limit is $ld\ n$. In this case, the distribution for all y_i of x is uniform, that is, for every instance of x, the probability p_i is $1/n$. That means, the more values a variable can take and the more equiprobable these values are, the higher the entropy value is.

In using this concept in order to calculate the entropy 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.

Following the procedure described by Smith and Silverberg (in revision), we calculated entropy values for word-initial letter–phoneme mappings in the five languages described above. The choice of languages was not typologically motivated but was dictated primarily by the available sources. For the present cross-linguistic research we needed corpora with a phonological transcription, based on the canonical pronunciation of words in isolation, or, alternatively, corpora of languages with a relatively regular orthography, so that a semi-automatic coding of their phonology was feasible.

We created the corpora for our cross-linguistic calculations by extracting lists from all mono- and polysyllabic word forms in the five languages

concerned, together with their pronunciations. For the three Germanic languages we used the Celex database (Baayen et al., 1995). The size of these corpora was about 66,594 word forms for English, 295,047 for Dutch, and 311,402 for German. Our French source was the Lexique database, Version 2 (New, Pallier, Brysbaert, & Ferrand, 2004), consisting of 128,908 word forms. To build the Hungarian corpus we took the Multext corpus (Erjavec, 2001), consisting of 62,384 word forms. As the Hungarian corpus, probably due to its very predictable pronunciation, did not contain phonological information, we coded the pronunciation ourselves and let a native speaker of Hungarian check the generated pronunciations. As a final step we extracted the word-initial letters/phonemes from the word forms in our corpora and calculated the entropy values for the different letter–phoneme correspondences.[1]

To give an example from our English corpus, words beginning with the letter ⟨b⟩, are always pronounced with the phoneme /b/, whereas words beginning with the letter ⟨w⟩ may be pronounced with four different phonemes, /w/, /r/, /h/, and /v/. In contrast, in Dutch both initial letters, ⟨b⟩ and ⟨w⟩, have only one possible pronunciation.

The result of calculating the entropy values for these mappings is presented in Table 1.

For each of the four possible pronunciations of the letter ⟨w⟩ the probability for the mapping was computed. In our case, among the 1697 words in

Table 1. Sample calculations for word-initial letter-to-phoneme entropy values

English				
1st letter	1st phoneme	example	# of occurrences	entropy components
b →	/b/	e.g. ***b**all*	3862	0.0
entropy value **H(b) = 0.0**				
w →	/w/	e.g. ***w**ell*	1549	–0.120168
→	/r/	e.g. ***w**rite*	121	–0.271655
→	/h/	e.g. ***wh**ole*	25	–0.089642
→	/v/	***w**eltanschauung*	2	–0.011466
entropy value **H(w) = 0.492931**				
Dutch				
1st letter	1st phoneme	example	# of occurrences	entropy components
b →	/b/	e.g. ***b**all*	22422	0.0
entropy value **H(b) = 0.0**				
w →	/ʋ/	e.g. ***w**ater*	10367	0.0
entropy value **H(w) = 0.0**				

the English corpus starting with ⟨w⟩, there were 1549 occurrences in which the word was pronounced with an initial /w/, amounting to a probability **p** of 0.9128. The probabilities for the other mappings were 0.0712 for /r/, 0.0147 for /h/, and 0.0011 for /v/, respectively.

Inserting the probabilities into the entropy formula, we multiplied each probability with the binary logarithm (**ld** or $\mathbf{log_2}$) of this probability. For the mapping ⟨w⟩–/w/ this was $0.9128 \cdot \log_2 0.9128 = 0.9128 \cdot -0.13163 = -0.120168$; the mapping ⟨w⟩–/r/ resulted in $0.0712 \cdot \log_2 0.0712 = -0.271655$; the mapping ⟨w⟩–/h/ yielded $0.0147 \cdot \log_2 0.0147 = -0.089642$. Finally, the mapping ⟨w⟩–/v/ generated $0.0011 \cdot \log_2 0.0011 = -0.011466$.

Finally we calculated the negative sum, yielding the entropy value $\mathbf{H(w)} = 0.492931$, describing the lack of predictability of the pronunciation of the word-initial letter ⟨w⟩ or, in other words, the degree of ambiguity of the variable ⟨w⟩.

As pointed out before, the higher the entropy value is, 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 0.492931, which results from the fact that ⟨w⟩ has one truly dominant pronunciation and three less frequent ones. If a letter in the English alphabet mapped with equal probability, completely unpredictably, to all the 46 phonemes in English, its (maximal) entropy value would be $\log_2 46$, that is 5.5236.

In contrast to the ambiguous pronunciation of English word-initial ⟨w⟩, the pronunciations of English word-initial ⟨b⟩ and Dutch word-initial ⟨b⟩ and ⟨w⟩ are completely predictable, so the entropy values for these letters are zero. Here, the equation is rather trivial, as in case of H(b) with a probability of $\mathbf{p} = 1$, $H(b) = 1 \cdot \log_2 1 = 1 \cdot 0 = 0$, as $\log_2 1 = 0 \leftrightarrow 2^0 = 1$.

Note that our method does not involve any kind of one letter to one phoneme alignment, that is usually done in research on or development of text-to-speech systems (e.g., Damper, Marchand, Anderson, & Gustafson, 1999). This alternative approach would require, at least for some languages, the assumption of *null phonemes* (that silent letters like ⟨w⟩ in *write* could map to), and *pseudo phonemes* or *null graphemes* (for phonemes corresponding to more than one letter, like ⟨x⟩ in *exact*). Whereas these null-correspondences might at first sight seem to be a more naturalistic description of the mappings, compared to the often skewed letter–phoneme mappings presented here, the psycholinguistic reality of null phonemes or null graphemes is highly debatable.

Table 2. Sample calculations for word-initial phoneme-to-letter entropy values

English				
1st phoneme	1st letter	example	# of occurrences	entropy components
/b/ →	*b*	e.g. ***b**all*	3862	0.0
entropy value **H(/b/) = 0.0**				
/k/ →	*c*	e.g. ***c**ar*	5578	–0.151360
→	*k*	e.g. ***k**ey*	348	–0.231349
→	*q*	e.g. ***q**uiche*	350	–0.232217
→	*x*	**x**mas	1	–0.002010
entropy value **H(/k/) = 0.616935**				
German				
1st phoneme	1st letter	example	# of occurrences	entropy components
/b/ →	*b*	e.g. ***b**all*	179127	0.0
entropy value **H(/b/) = 0.0**				
/f/ →	*v*	e.g. ***v**ielseitig*	17032	–0.440697
→	*f*	e.g. ***f**reundlich*	10432	–0.530736
→	*p*	e.g. ***p**honogisch*	824	–0.148599
entropy value **H(/f/) = 1.120032**				

To gain an insight into the (*feedback*) sound-to-spelling patterns as well, we analogously calculated the entropy values of the word-initial phoneme–letter mappings. For example, English words that are pronounced word-initially with the phoneme /k/ can start with the letters ⟨c⟩, ⟨k⟩, or ⟨q⟩. A word with initial /b/, however, is always spelled with a ⟨b⟩. In German, words pronounced with an initial /f/ can be written with initial letters ⟨f⟩, ⟨p⟩, or ⟨v⟩, whereas words with initial /b/ are always starting with the letter ⟨b⟩ (see Table 2).

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

4. Results and Discussion

With respect to word-initial single letter to single phoneme and single phoneme to single letter entropies, the position of the five languages examined is shown in the following two-dimensional entropy table.

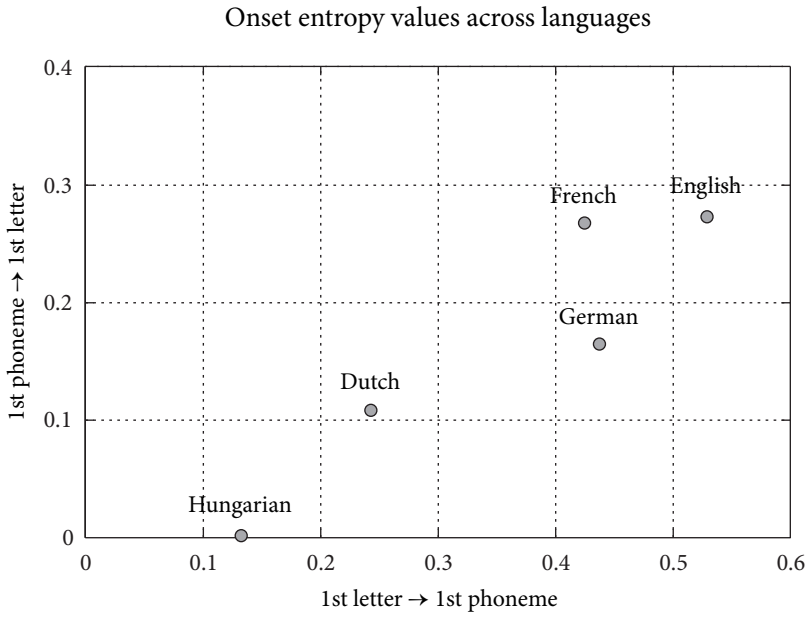

Figure 1. Deviations from a 1 : 1 mapping between word-initial letters and phonemes stated in entropy values. The values on the x-axis show the degree of spelling-to-sound ambiguity, whereas the y-axis depicts the degree of sound-to-spelling ambiguity.

If a hypothetical orthography were fully transparent at the letter–phoneme level, that is, if any letter would always map onto the same phoneme and vice versa, its entropy values in Figure 1 would fall onto the origin of the system of co-ordinates, that is, they would equal (0,0). No orthography investigated here is that transparent. The higher the position on the x-axis, the more ambiguous the language is in (feedforward) spelling-to-sound direction. The higher the position on the y-axis, the more ambiguous the language is in (feedback) sound-to-spelling direction.

The results show clearly that when comparing word-initial letter-to-phoneme mappings, English has the most ambiguous orthography, followed by, in descending order, German, French, and Dutch, with Hungarian having the most predictable orthography of the five languages examined.[2]

Also in the reverse direction, that is, in terms of phoneme-to-letter mappings, it is English that has the most ambiguous orthography, closely followed by, in descending order, French and then German, Dutch, and Hungarian.

In comparing both values, a general shift towards higher values on the x-axis can be noticed. The values on the x-axis are always higher than the

corresponding values on the y-axis. This can be interpreted as a consequence of the general asymmetry between the number of letters and phonemes that exists in most languages. All languages investigated in this study have a larger phoneme inventory than letters in their alphabet to denote these phonemes. Therefore languages often use *graphemes* composed of several letters, that is, *digraphs* or *trigraphs,* to express a single phoneme, for example, ⟨ou⟩ in *out* or ⟨Sch⟩ in *Schiff*. When looking at the pronunciation of a single letter as compared to the spelling of a single phoneme, this leads to greater ambiguities in the letter-to-phoneme direction than in the phoneme-to-letter direction.

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 (Kessler, personal communication, November 2002).

High letter-to-phoneme entropy values can reflect letter–grapheme complexities (e.g., the English ⟨w⟩, that could be part of the grapheme ⟨wr⟩ as in *write*). This constitutes the first component of the overall ambiguity of first letter to first phoneme mappings. Even Hungarian, that generally shows very reliable grapheme–phoneme mappings, displays a certain amount of ambiguity on the letter–phoneme level due to the fact that some letters are pronounced differently in isolation than as part of a grapheme.

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, that is, the fact that some ambiguous pronunciations can be resolved by taking the subsequent letters into account. In other words, context beyond the grapheme level may disambiguate ambiguous letter-to-sound relations. An example is the initial ⟨c⟩ in English and Dutch, whose pronunciation is quite reliably determined by the place of articulation of 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 ⟨v⟩ in German, that is pronounced like /f/ in *Vater*, but like /v/ in *Vase*, that is, an inherent ambiguity that cannot be resolved in a larger context.

In looking only at the mappings between first letters and first phonemes, we cannot decide which of these three components (that might influence reading performance to different degrees, see Martensen et al., 2003) contributes to what degree to the overall entropy value. The entropy value concerning first letter to first phoneme mappings expresses these different types of ambiguity in one single number. In order to disentangle these three components, we proceeded by analyzing mappings that take larger grain sizes into account.

Table 3. Sample entropy calculations for the mappings between the first two letters and the word-initial phoneme.

English				
1st 2 letters	1st phoneme	example	# of occurrences	entropy components
sh →	/ʃ/	e.g. ***sh**ip*	693	0.0
entropy value: **H(sh) = 0.0**				
th →	/θ/	e.g. ***th**ick*	398	–0.125545
→	/ð/	e.g. ***th**is*	35	–0.291307
→	/t/	e.g. ***th**yme*	5	–0.073663
entropy value: **H(th) = 0.490515**				
German				
1st 2 letters	1st phoneme	example	# of occurrences	entropy components
ph →	/f/	e.g. ***ph**onetik*	842	0.0
entropy value: **H(ph) = 0.0**				
ch →	/k/	e.g. ***ch**aos*	181	–0.391025
→	/ç/	e.g. ***ch**emie*	45	–0.429421
→	/ʃ/	e.g. ***ch**ance*	33	–0.369196
→	/tʃ/	e.g. ***ch**arter*	13	–0.209674
entropy value: **H(ch) = 1.399315**				

Our next step was to distinguish between truly unpredictable letter–phoneme mappings (that cannot be disambiguated by a larger context) on the one hand and mere ambiguities on the letter–phoneme level that are disambiguated in a larger context on the other hand. We thus calculated entropy values for larger units such as the mappings between the first two or three letters and the first phoneme. Example calculations for entropy values for mappings between the first phoneme and the first two letters are presented in Table 3.

Analogously, we calculated phoneme-to-letter correspondences for different grain sizes such as the mappings between the first two or three phonemes and the first letter. Examples are reported in Table 4.

In doing so for all the words in our five languages, we can investigate how the number of ambiguities that are resolved increases by increasing the grain size that is taken into account. Figure 2 shows the overall word-initial entropy values for different grain sizes. The x-axis shows the degree of spelling-to-sound ambiguity, the y-axis the degree of sound-to-spelling ambiguity, for grain sizes of *one* (black), *two* (gray) and *three* (white) letters/phonemes, respectively.

Table 4. Sample entropy calculations for the mappings between the first two phonemes and the word-initial letter.

English				
1st 2 phonemes	1st letter	example	# of occurrences	entropy components
/ʃʌ/ →	*s*	e.g. *shuffle*	51	0.0
entropy value: **H(/ʃʌ/) = 0.0**				
/ʃiː/ →	*c*	e.g. *chic*	4	–0.363231
→	*s*	e.g. *she*	30	–0.159328
entropy value: **H(/ʃiː/) = 0.522559**				
Dutch				
1st 2 phonemes	1st letter	example	# of occurrences	entropy components
/sf/ →	*s*	e.g. *sfeer*	37	0.0
entropy value: **H(/sf/) = 0.0**				
/ʃiː/ →	*c*	e.g. *chirurg*	57	–0.266819
→	*s*	e.g. *shiitake*	15	–0.471466
entropy value: **H(/ʃiː/) = 0.738285**				

These expanded computations show that when taking larger contexts into account, the pattern of relative ambiguity of the five languages as compared to one another remains about the same.[3]

At a grain size of three letters, English remains more ambiguous than the other orthographies in the spelling-to-sound direction, whereas French, Dutch, and German approximate one another, and Hungarian does not display any ambiguities anymore.

Also in the opposite, sound-to-spelling, direction, the relative ambiguity pattern remains largely the same. The drop in ambiguity by increasing the grain size is not as large as in the spelling-to-sound analyses, as the first component (asymmetry in letter–phoneme mappings) cannot contribute to the entropy values in the reverse direction. Except for few letters like for example the letter ⟨x⟩, that can denote two phonemes with one letter, there are not many cases where one single letter maps to more than one phoneme (whereas in the spelling-to-sound direction quite often two or three letters form a grapheme and as such map to a single phoneme). However, in the sound-to-spelling direction, the disambiguating factor when taking larger grain sizes into account can partly be attributed to *orthographic constraints*. For example, when hearing word-initial /k/ in English, we cannot be sure whether the word in question is written with ⟨c⟩, ⟨k⟩, or ⟨q⟩. If the onset unfolds to the cluster /kl/, we know

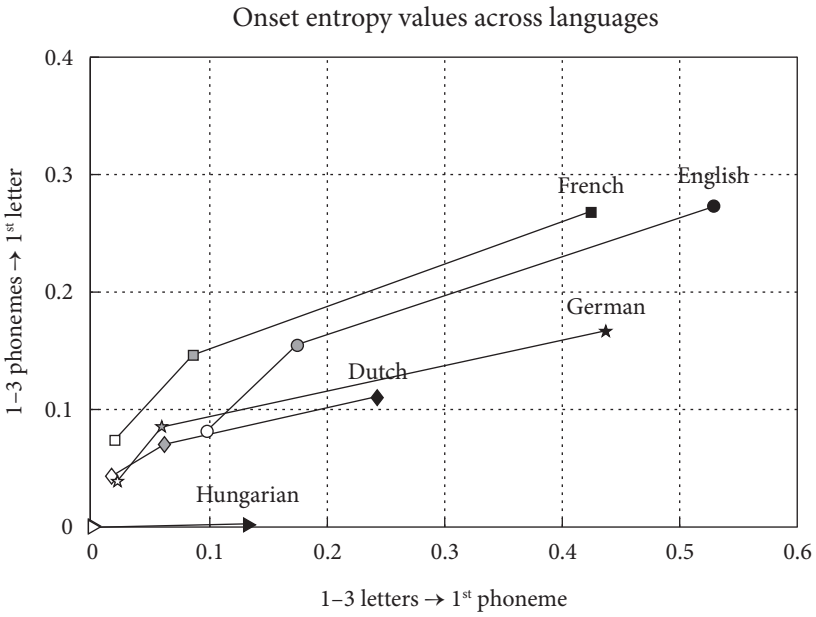

Figure 2. Decreasing ambiguity by increasing word-initial orthographic and phonological contexts.

that it is written with either ⟨c⟩, or ⟨k⟩, as the onset cluster *⟨ql⟩ does not exist in English.

The calculations above, where we gradually expanded the grain size of analysis might raise the question why we chose to first calculate letter–phoneme mappings and only later acknowledge the existence of graphemes, instead of starting with graphemes. Graphemes might be considered better candidate units for an investigation into a language's orthography than letters, as languages' orthographies are built on grapheme–phoneme correspondences.

The reason for not doing so is that, even though at first sight it seems to be a rather straightforward task to generate a set of grapheme–phoneme-correspondence rules, it is not that easy at all to identify the graphemes themselves, especially in the case of relatively opaque orthographies. For example, there is quite a variation in the number of graphemes and the grapheme–phoneme alignment proposed for English. Whereas the number of existing English phonemes varies only slightly across the different analyses of the sounds of English, depending on the variety analyzed (e.g., 15 vowel phonemes for American English, and 20 vowel phonemes for British English, Ladefoged, 2001) or on the status of some marginal loan word phonemes, the number of corresponding

graphemes varies substantially. Reported sizes of the English grapheme inventory range from 52 (Lesetar, 1997), to 108 (Wijk, 1966), 110 (Carney, 1994), and 195 (Gontijo, Gontijo, & Shillcock, 2003). Consequently, also quite different sets of grapheme–phoneme correspondence rules are bound to exist, as the grapheme–phoneme alignment itself is not a trivial procedure and sometimes arbitrary. For example, words with silent letters like ⟨gh⟩ in *light* or ⟨u⟩ in *guide* can be segmented into different graphemes and subsequently lead to different grapheme–phoneme mappings (cf. Andrews & Scarratt, 1998; Kessler & Treiman, 2001). In contrast, the letter unit is clearly defined and less arbitrary. However, the proposed word-initial entropy calculations can in principle be adapted to operate on other sublexical units than letters or phonemes such as for example on graphemes. In this way, mappings involving silent letters, for example, would yield much smaller entropy values, indicating less ambiguity.

5. Conclusions

In this paper we investigated ambiguities in word-initial bi-directional letter–phoneme mappings in five languages. The orthographies examined deviate to various degrees from an "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 et al., 1995; Ziegler et al., 1997) we found that in terms of spelling-to-sound and sound-to-spelling relations at the word-initial letter–phoneme level, English has the most opaque, ambiguous orthography, and Hungarian the most transparent, unambiguous one. The other three languages lie in between; the French and German orthographies are relatively ambiguous, although not as much as English, whereas the Dutch orthography is best characterized as intermediate in terms of spelling–sound transparency. This pattern remains about the same when mappings of a larger grain size, namely, between the first two or three letters and the first phoneme, are taken into account.

The word-initial entropy paradigm presents an alternative to standard methods to assess languages' orthographic transparency (e.g. the rime-consistency paradigm) and has in our view distinct advantages over the traditional approaches.

The entropy paradigm is relatively simple to compute, and language-independent. It requires only basic information, namely, a list of words together with their pronunciation, and does not need any other information about the

language properties, as for example, morphological and syllable properties, or sublexical and graphemic structures.

Furthermore, the entropy paradigm allows an analysis of virtually all the languages' words, not only the monomorphemic and monosyllabic subset that might be very small in morphologically rich languages, thereby increasing the sample size considerable.

A main advantage of using the entropy concept in order to analyse languages' spelling–sound relations is that it allows for a multitude of variations. The method can easily be altered in order to, for example, vary the grain size of analysis or the position of the spelling/pronunciation pattern to be examined. Future research using word-initial entropy might proceed to investigate the (un)ambiguity of other unit sizes. A possible extension towards the grapheme level would allow the calculation of grapheme–phoneme entropy values for all the words' graphemes and phonemes, covering the whole words, and not limiting the analyses to the onset.

In summary, the research presented here leads us to the conclusion that word-initial entropy values provide a valid basis for assessments of the transparency of orthographies.

In addition to merely computing languages' overall spelling–sound ambiguity patterns, future research might explore the empirical validity of the computational analyses (e.g., Borgwaldt, Hellwig, & De Groot, in revision). As spelling–sound ambiguities are known to affect reading performance, an investigation into the psycholinguistic relevance of the spelling–sound (un)ambiguities found might indicate whether irregularities at different grain sizes result in longer reaction times in visual word recognition tasks. In other words, do the computed metrics of spelling–sound unpredictability affect the actual processing? Are words with higher onset entropy processed more slow than words with lower onset entropy? The impact of onset entropy on language processing might depend on (type of) language, processing task, and/or the direction of ambiguity, and a detailed investigation into the impact of spelling–sound ambiguities on reading performance is a necessary complement and expansion in order to test the psycholinguistic applicability of the generated entropy results presented above.

Notes

1. For the sake of simplicity we will use the term *words* instead of *word forms* in the remainder of the paper.

2. As the corpora for our languages were of different size — the German corpus is roughly five times larger than the Hungarian corpus — we tested for possible confoundings of overall entropy values and corpus size. We randomly extracted from each language corpus a list of 50,000 word forms and calculated entropy values for these subsets. The general pattern stayed the same, leading to the conclusion that the present entropy values reflect the language's degree of orthographic (un)ambiguity quite truthfully.

3. As a reviewer pointed out, the drop in average onset entropy when taking larger context into account might be exaggerated by a relatively large number of zero values resulting from transparent mappings. For example, when computing the mappings between the first two letters and the first phoneme, all transparent mappings, e.g. ⟨ta⟩–/t/, (*table*), ⟨te⟩–/t/ (*test*), ⟨ti⟩–/t/ (*tile*), ⟨to⟩–/t/ (*tooth*), ⟨tr⟩–/t/ (*transport*) etc. contribute separately to the average entropy value.

References

Abondolo, D. 1998. The Uralic languages. London: Routledge.

Andrews, S., & Scaratt, D. 1998. Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds? Journal of Experimental Psychology: Human Perception and Performance 24(4):1052–1086.

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

Berndt, R. S., Reggia, J. A., & Mitchum, C. C. 1987. Empirically derived probabilities for grapheme-to-phoneme correspondences in English. Behavior Research Methods, Instruments, and Computers 19(1):1–9.

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

Borgwaldt, S. R., Hellwig, F. M., & De Groot, A. M. B. In revision. Onset entropy matters — Letter-to-phoneme mappings in seven languages. Ms. University of Amsterdam.

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

Carney, E. 1997. English spelling. London: Routledge.

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

Coltheart, M. 2003. Implementation of the French DRC. Unpublished Manuscript. [Online] http://www.maccs.mq.edu.au/~max/DRC/FrenchDRC.doc.

Damper, R., Marchand, Y., Anderson, M., & Gustafson, K. 1999. Evaluating the pronunciation component of text-to-speech systems for English: A performance comparison of different approaches. Computer Speech and Language 13(2):155–176.

Dewey, G. 1971. English spelling: Roadblock to reading. New York: Teachers College Press.

Eisenberg, P. 1998. Grundriß der deutschen Grammatik. Das Wort. Stuttgart: Metzler.

Erjavec, T. 2001. Harmonised morphosyntactic tagging for seven languages and Orwell's 1984, 6th Natural Language Processing Pacific Rim Symposium, Tokyo, NLPRS'01, 487–492.[Online] http://nl.ijs.si/ME/.

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

Gontijo, P. F. D., Gontijo, I., & Shillcock, R. 2003. Grapheme–phoneme probabilities for British English. Behavior Research Methods, Instruments, & Computers 35(1):136–157.

Gottlieb, H. 2001. Hvilke stavefejl er rigtige? Om sikre sprogbrugere og det usikre sprog. Retorikmagasinet 11(41):9–11.

Harris, M., & Vincent, N. (eds.). 1990. The Romance languages. London: Routledge.

Harris, M., & Giannouli, V. 1999. Learning to read and spell in Greek: The importance of letter knowledge and morphological awareness. In Learning to read and write: A cross-linguistic perspective, M. Harris and G. Hatano (eds.), 51–70. Cambridge: Cambridge University Press.

Karlsson, F. 1983. Suomen kielen äänne- ja muotorakenne. Juva: WSOY.

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

König, E., & van der Auwera, J. 2003. The Germanic languages. London: Routledge.

Ladefoged, P. 2001. Vowels and consonants. Oxford: Blackwell.

Lange, M., & Content, A. 1999. The grapho-phonological system of written French: Statistical analysis and empirical validation. Paper submitted to the 37th Annual Meeting of the Association of Computational Linguists. [Online] http://acl.ldc.upenn.edu/P/P99/P99–1056.pdf.

Lesetar, P. 1997. Say it right! English pronunciation dictionary. Doctoral Dissertation, University of Alberta.

Maraschio, N. 1993. Grafia e ortografia: evoluzione e codifiazione. In Storia della lingua italiana, Vol. I, L. Seriami and P. Trifone (eds.) , 139–227. Turin: Guilio Einaudi Editore.

Martensen, H., Maris, E., & Dijkstra, T. 2000. When does inconsistency hurt? On the relation between consistency effects and reliability. Memory & Cognition 28(4):648–656.

Martensen, H., Maris, E., & Dijkstra, T. 2003. Phonological ambiguity and context sensivity. On sublexical clustering in visual word recognition. Journal of Memory and Language 49 (3):375–395.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. 2004. Lexique 2: A New French Lexical Database. Behavior Research Methods, Instruments, and Computers 36(3):516–624.

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

Posner, R. 1996. The Romance languages. Cambridge: Cambridge University Press.

Seymour, P. H. K., Aro, M., & Erskine, J. M. 2003. Foundation literary acquisition in European orthographies. British Journal of Psychology 94(2):143–174.

Shannon, C. E. 1948. A mathematical theory of communication. Bell System Technical 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.

Vago, R. M. 1980. The sound pattern of Hungarian. Washington DC: Georgetown University Press.

Venezky, R. L. 1970. The structure of English orthography. The Hague: Mouton.

Venezky, R. L. 1999. The American way of spelling. New York: Guilford.

Véronis, J. 1986. Etude quantitative sur le système graphique et phonologique de français. Cahier de Psychologie Cognitive 6:501–531.

Wijk, A. 1966. Rules of pronunciation for the English language. 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.

Authors' addresses:

Susanne R. Borgwaldt
University of Sydney
School of Psychology
NSW2006, Australia

e-mail: suske7@yahoo.com

Frauke M. Hellwig
Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen
The Netherlands

e-mail: Frauke.Hellwig@mpi.nl

Annette M.B. de Groot
University of Amsterdam
Expertisecentrum Academische Zaken
Spui 21
1012 WX Amsterdam
The Netherlands

e-mail: adegroot@bdv.vva.bl