Onset Entropy in Language(s) - Comparing Alphabetic Orthographies

S.R. Borgwaldt[1], F.M. Hellwig[2], A.M.B. de Groot[1]

[1]University of Amsterdam

[2]Max Planck Institute for Psycholinguistics & F.C. Donders Centre for Cognitive Neuroimaging

Over the past several decades a considerable number of studies have investigated the role of spelling-sound transparency in visual word recognition. Alphabetic languages show more or less ambiguous relations between spelling and sound patterns. In languages with a transparent orthography, like Hungarian or Italian, the pronunciation can be predicted from the spelling and vice versa. In languages with an opaque orthography, like English, spelling-pronunciation correspondences are often quite unpredictable and ambiguous.

Recent investigations into the (ir)regularity of alphabetic orthographies (cf. Treiman et al. [1] for English, Lange & Content [2] for French, and Martensen et al. [3] for Dutch) have focussed on expressing this ambiguity in terms of entropy values, a concept proposed by Shannon [4].

Entropy is a measure of the ambiguity of a variable and is defined as follows: for a variable **x,** which can take **n** values (**y_1, y_2, ... y_n**) with the probability of **p**, the entropy of **x**, **H(x)** is:

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

If a variable takes only one value, its entropy equals 0.
The more values a variable can take, and the more similar the probabilities for these values are, the higher its entropy value is.

Method

Following the procedure described by Smith & Silverberg [5], we calculated entropy values for bi-directional letter-phoneme onset mappings in six languages (Dutch, (British) English, French, German, Hungarian and Italian). This computational analysis allowed us to position each of these languages in relation to the others on the continuum from deep to shallow orthographies, with respect to both spelling-to-sound and sound-to-spelling ambiguity.

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" could be pronounced with four different phonemes, /w/, /r/, /h/, /v/. Calculating the entropy values for these mappings results in:

1st letter	1st phoneme		# of occurrences	entropy value
b →	/b/	e.g. *ball*	2393	0.0
			H(b) =	**0.0**
w →	/w/	e.g. *well*	1005	0.113565
→	/r/	e.g. *write*	67	0.246623
→	/h/	e.g. *whole*	22	0.113260
→	/v/	*weltanschauung*	1	0.009221
			H(w) =	**0.482669**

To gain an insight into the (feedback) sound-spelling patterns as well, we analogously calculated the entropy of the word-initial phoneme-letter mappings. For example, in English, words that are pronounced with the phoneme /k/ in their onset can start with the letters "c", "k", or "q". A word with initial /b/, however, is always spelled with a "b".

1st phoneme	1st letter		# of occurrences	entropy value
/b/ →	b	e.g. *ball*	2393	0.0
			H(/b/) =	**0.0**
/k/ →	c	e.g. *car*	3386	0.156571
→	k	e.g. *key*	221	0.237539
→	q	e.g. *quiche*	221	0.237539
			H(/k/) =	**0.631648**

To calculate the overall onset entropy value of a language, we summed the entropy values for all single letters/phonemes (vowels & consonants) weighted by their frequency of occurrence within the corpus, i.e. with the probability, of a word beginning with this letter/phoneme.

Results

With respect to word-initial 1st letter to 1st phoneme and 1st phoneme to 1st letter entropies, the position of the six languages examined is shown in the following two-dimensional entropy figure. A completely predictable language would be positioned at (0,0), the origin of the coordinate plane.

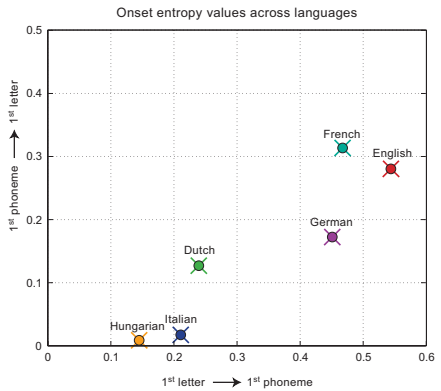

Figure 1: Deviations from a 1:1 mapping between word-initial letters and phonemes stated in entropy values. The values on the x-axis show the degree of spelling-sound ambiguity, whereas the y-axis depicts the degree of sound-spelling ambiguity.

In order to distinguish between truly unpredictable letter-phoneme mappings on the one hand and ambiguities caused by regular di- and trigraphs (like English "sh-ip" or German "Sch-iff") on the other hand, we additionally calculated entropy values for larger units such as the mappings between the first two or three letters and the 1st phoneme. Analogous calculations for phoneme-letter correspondences were carried out as well.

1st 2 letters	1st phoneme		# of occurrences	entropy value
sh →	/ʃ/	e.g. *ship*	409	0.0
			H(sh) =	**0.0**
th →	/θ/	e.g. *thick*	265	0.206575
→	/ð/	e.g. *this*	44	0.397288
→	/t/	e.g. *thyme*	5	0.093107
			H(th) =	**0.698969**

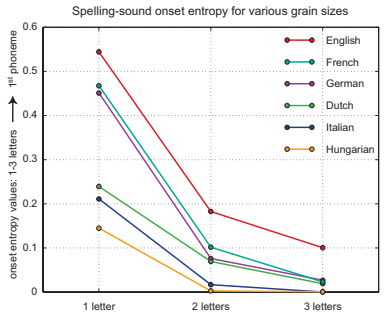

Figure 2: Decreasing pronunciation ambiguity by increasing orthographic onset window.

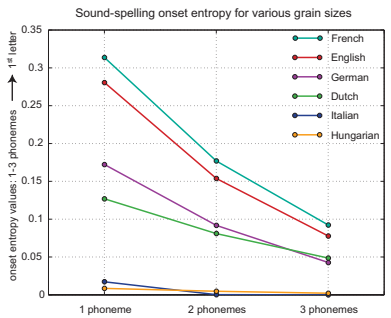

Figure 3: Decreasing spelling ambiguity by increasing phonological onset window.

Empirical Validation

We correlated the onset letter-phoneme entropy values with the word naming latencies of large scale studies carried out in three of the six languages, Italian (Barca et al. [6]), Dutch, and English (de Groot et al. [7]), controlling for onset cluster and voice key effects. For all three languages of varying orthographic transparency the correlations were highly significant: the higher the onset entropy, the longer the reaction times.

language	partial correlation coefficient (letter-phoneme entropy – RTs)	p-value (two-tailed)
Italian	0.1806	< 0.001
Dutch	0.1706	< 0.001
English	0.2424	< 0.001

Conclusions

- All orthographies examined deviate to various degrees from the "ideal" 1:1 mapping between letters and phonemes.
- Even in very regular languages these deviations influence reaction times in visual word recognition tasks.
- Onset entropy values provide a sound basis for assessments of orthographic transparency.

References

[1] Treiman, R., Mullenix, 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.

[2] 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 for Computational Linguists.

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

[4] Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, July, 379-423, & October, 623-656.

[5] Smith, J. & Silverberg, N. (in revision). Frequency and correspondence of initial letter/phoneme combinations for English words.

[6] 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.

[7] 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.

Correspondence address:
Susanne Borgwaldt
Department of Psychology
University of Amsterdam
Roetersstraat 15
1018 WB Amsterdam
e-mail: sborgwaldt@fmg.uva.nl