Write as you speak? A cross-linguistic investigation of orthographic transparency in 16 Germanic, Romance and Slavic languages

Anja Schüppert[1], Wilbert Heeringa[2], Jelena Golubovic[1] & Charlotte Gooskens[1]

[1]University of Groningen, [2]Fryske Akademy

1. Introduction

Ever since the invention of the printing press, the creation of a written standard represented an important cornerstone in the development of most languages. For many language varieties, this meant taking the step from a dialect or regiolect to a language, and for most national languages, the presence of a written standard was a necessary prerequisite. The orthographies of most European languages were developed from a set of unstandardized conventions which usually served as the basis of the new norm. For some languages the spelling norms were established with the first official translation of the Bible (e.g. Czech), while others set the norm through publishing a dictionary and describing the new spelling used (Dutch, French, Spanish). The spelling was then updated through a series of reforms, most of them passed in the 18th and 19th centuries. These reforms often followed significant historical changes and were seen as a vital part of language standardisation, which at the time was often an important element in nation building.

Naturally, since the speed at which the orthographic reforms follow the changes in speech varies substantially per language, this generally means that the extent to which ancient pronunciation is preserved in the current spelling also varies. Sometimes spelling is also deliberately kept unchanged in order to illustrate word etymology. This variety in decisions and the very fact that spelling rules are arbitrary, in turn leads to cross-linguistic differences in how accurately contemporary pronunciation is reflected in the orthography.

All official EU languages are written in alphabetic systems, which use roughly one character (or grapheme) for one sound (or phoneme). However, the correspondence of one character to one sound is not always a strict rule. This is partly because pronunciation has changed since the establishment of a written standard and the spelling has not been (fully) reformed to reflect these changes. Another reason is the fact that in many alphabetical orthographies, the grapheme repertoire is restricted to some 20-40 graphemes. The sound repertoires, however, are usually somewhat larger, which means that not every grapheme can be assigned to exactly one phoneme.

More specifically, the same grapheme can represent different phonemes, such as the letter <o> that can represent /ɔ/ as in <not>, or /oʊ/ as in <go>, or even /ɪ/ as in <women> in English. Analogically, the same phoneme can be represented by different graphemes, such as the consonant /i:/ that can be spelled <e> as in <here>, or <ea> as in <beat>, or <ee> as in <beet>. Similar examples can be found in other European languages, such as the pronunciation of the Dutch letter <a> as /ɑ/ in <pad>, but as /a:/ as in <paden>, or the spelling of the Swedish consonant /ɧ/ as <sk> in the word <skär>, or as <skj> in the word <skjorta>, or as <stj> in the word <stjärna>, or as <g> in the word <giraf>.

Not only can the same phoneme be represented by different graphemes and vice versa, but in many European orthographies, a single phoneme might also be represented by a grapheme *cluster* (such as the clusters /ph/ or /ea/ in English e.g. <phonetic measurement> /fə'netɪk meʒəmənt/). To be able to read the written word <phonetic> correctly, the reader needs to be familiar with the rule that the grapheme cluster /ph/ is not pronounced as the sum of its parts /ph/ or /p^h/, but as a separate phoneme /f/. The same is true for the cluster <ea> in <measurement>, which is generally transcribed as /e/ (British English) or /ɛ/ (American English). There are numerous other examples from different European languages, such as the grapheme cluster <sch> in German, which is pronounced /ʃ/ in <syntaktisch> /zʏn'taktɪʃ/ or the clusters <gn> and <ent> in French <alignement> /aliɲmɑ̃/ which are pronounced with the single phonemes /ɲ/ and /ɑ̃/, or the spelling <Groningen> with the grapheme cluster <ng>, for the pronunciation with a single phoneme /ŋ/. In other words, a reader who is aware of the fact that a single letter can be pronounced in different ways, but not of these additional rules for the pronunciation of grapheme clusters, will have even more difficulties to read written language properly – although arguably, the degree to which these missing rules create problems in the reading process vary a lot across different languages and their orthographies. And again, in a similar way, phoneme clusters can be spelled with a single grapheme, such as the affricate /dʒ/ in the English name /dʒɒn/ (Am. /dʒɑn/) which is spelled <John>, or the phoneme cluster /ks/ in /daɪəlekt prɒksɪməti/ which is spelled <dialect proximity>. These intransparencies create huge problems for beginning readers and writers, be it (usually young) native speakers or (young as well as older) second-language learners.

And as if the sheer presence of rules for grapheme (and phoneme) clusters was not enough to confuse the poor reader (and writer), in some orthographies similar grapheme clusters have different etymological backgrounds. This means that they follow different pronunciation rules and thus are not clearly decipherable, even if a reader is aware of the fact that grapheme clusters ought to be treated differently than single graphemes. One of many examples for this is the English grapheme cluster <gh> which can be pronounced as /f/ (as in <laugh> /lɑ:f/), or as a zero-ending as in <dough> /doʊ/. To be able to read words such as <laugh> and <dough> correctly, the reader needs to apply not only the knowledge that <gh> is pronounced differently than the sum of its parts, but also be familiar with the etymology of the word in order to choose the correct pronunciation (alternatively, have learned proper pronunciation for every specific word separately, which might be the most frequent strategy). Analogically, again, if every /f/ sound in English would be spelled /gh/, this would restrict the opaqueness and hence the difficulties to the process of reading. But what makes things even worse is the fact that the phoneme /f/ can also be spelled /f/ or /ff/, and that the phoneme cluster /oʊ/ also can be spelled <o> (as in <go>), <ow> (as in <low>), or <oe> (as in <toe>). This extends the scope of intransparency from reading (due to an opaque grapheme-phoneme correspondence) to writing (due to an opaque phoneme- grapheme correspondence).

If you are able to read this paper until here, you are most likely at least partly familiar with English spelling. You might even have come across the notion of English orthography belonging to the orthographies that are particularly intransparent - an assumption which strongly contradicts the statement made by Chomsky & Halle (1968: 49) who claimed that "[t]here is [...] nothing particularly surprising about the fact that conventional orthography is [...] a near optimal system for the lexical representation of English words". There is not only common-sense and speculative evidence that Chomsky & Halle (1968) were wrong, but also growing scientific support for the objection against their claim. Borgwaldt, Hellwig & De Groot (2005) conducted entropy (= uncertainty) measurements for letter-to-phoneme mappings in Dutch, English, French, German, Hungarian, Italian and Portuguese. They reported that English had the highest entropy value of all included languages, which means that the pronunciation predictability for English letters is lower than for any of the other six languages. Their algorithm models an advanced reader, in that it analyses the consistency of spelling and pronunciation of rimes and onsets, and not of single graphemes and phonemes. In particular, some basic knowledge of consonant and grapheme clusters is presumed, such as the

English rule that <gh> is only pronounced /f/ word-medially and word-finally, but never word-initially. In contrast to Borgwaldt et al.'s (2005) bottom-up approach, Nicolai & Kondrak (2015) used a top-down approach to scrutinise Chomsky & Halle's (1968) claim. Instead of quantifying the uncertainty of different existing orthographies, they developed an algorithm modelling a 'better fit' for the spelling of English pronunciation, i.e. a more transparent orthography for English. The two investigations have two things in common: The conclusion that English orthography is far from being optimal, and the fact that they restrict themselves to phoneme-to-grapheme correspondences (quantifying spelling problems in English), but do not investigate grapheme-to-phoneme correspondences (which would quantify reading problems).

It becomes evident that English orthography is situated close to one end of the continuum between transparent (or *shallow*) and intransparent (or *deep*) orthographies. Some of the European orthographies that are said to belong to the other end of this continuum are Finnish and several Slavic languages, in which most graphemes are pronounced in only one way, and most phonemes are spelled in only one way – at least in careful speech. According to the *Orthographic Depth Hypothesis* (ODH) (Katz & Frost 1992), decoding a deep orthography requires a different way of reading than decoding a shallow one. The ODH suggests that when reading a shallow orthography, the reader can focus on phonological (non-lexical) information, while the reader has to focus on larger units (lexical information) when reading a deep orthography. In line with the ODH, the *Psycholinguistic Grain Size Theory* (PGST), put forward by Ziegler & Goswami (2005), postulates that reading shallow orthographies allows the reader to process smaller units (*grain sizes*) than reading deep orthographies, since the predictability of phoneme-grapheme correspondences in deep orthographies increases if the grapheme or phoneme context is taken into account. In other words, readers and writers of deep orthographies are more likely to use logographic entities than learners of shallow orthographies are, at least when they have reached a certain level of literacy.

Seymour, Aro & Erskine (2003) were among the first to conduct a cross-linguistic study on literacy acquisition on a broad range of European languages. They use *syllable complexity* and *orthographic depth* as two independent predicting factors. Although the outcomes of such a study have to be interpreted with caution (as the orthographic system naturally is not the only factor that differs between 1st-graders in Scotland, Iceland, or Greece), their study showed that children acquiring an orthography that has been described as relatively deep (such as English) are learning to read twice as slowly as children acquiring an orthography that is traditionally seen as shallow (such as Finnish). However, the classification into 'deep' or 'shallow' (on a five-pointscale) is a "hypothetical classification" (Seymour et al 2003: 146). It is not clear what the basis of the placement of every orthography on the continuum between 'deep' and 'shallow' in their study is based on. Also earlier cross-linguistic studies investigating the effect of orthographic depth on reading development lay a cloak of silence on the question how orthographic depth was determined (cf. Wimmer & Goswami 1994, Frith, Wimmer & Landerl 1998, Goswami, Porpodas & Wheelwright 1997, Goswami 2008, Ziegler & Goswami 2005, Ziegler & Goswami 2006). However, in a study comparing children's reading speed and accuracy in three languages (English, French and Spanish), Goswami, Gombert & de Barrera (1998) refer to investigations on the consistency of English spelling conducted by Treiman, Mullennix, Bijeljac-Babic & Richmond-Welty (1995), as well as a similar investigation by Peereman & Content (1996) on French orthographic consistency. No measurements are presented for Spanish, however. Ziegler, Stone & Jacobs (1997) present a database of phoneme-to-grapheme and grapheme-to-phoneme consistency for 2,694 English monosyllabic words and report that 72.3% of the monosyllabic words could theoretically be spelt in more than one way, and that 30.7% of the words could be pronounced in more than one way. In a recent paper, Ziegler, Bertrand, Tóth, Csépe, Reis, Faísca, Saine, Lyytinen, Vaessen & Blomert (2010) conduct a thorough investigation of reading skills in Finnish-, Hungarian-, Dutch-, Portuguese-, and French-speaking children, and discuss their findings in the light of the mean orthographic depth of the word onsets in every language as established by Borgwaldt, Hellwig & De Groot (2005).

Crucially, however, most studies that compared cross-linguistic literacy acquisition in a large number of languages seem to categorise the involved orthographies on the basis of 'common sense' or speculation. In the present paper the orthographies of 16 European languages are compared using a uniform methodology applied to the same set of 100 words. Importantly, the method models a completely illiterate reader, as the entity of our analysis are phonemes and graphemes. In other words, any rules that are reflected in larger entities of written language, such as clusters or rimes, are treated as the sum of their parts. By providing entropy values for both grapheme-to-phoneme and phoneme-to-grapheme correspondences, the results can be used as a basis for the prediction of writing as well as reading development in very early beginners, namely practically illiterate beginners who are only familiar with the 'names' of the letters in their alphabet.

2. Method

Entropy measures the uncertainty in a random variable. In this study, the uncertainty between phoneme-to-grapheme mappings and grapheme-to-phoneme mappings are at focus: a letter may correspond to one or several phonemes. If the letter can correspond to more than one phoneme, then for a beginning reader, who knows nothing more than the fact that s/he is confronted with an alphabetical orthography as well as the names of the letters of the specific alphabet, there exists uncertainty about which phoneme corresponds to the specific letter in the word that s/he is reading. Reversely, when the same person is listening to a language in which the same phoneme in different words is transcribed by the different letters, there will be uncertainty about which letter represents the sound s/he hears when writing down the word.

We quantify uncertainty as Shannon's entropy (Shannon 1948). Given a grapheme *x*, and given variable *Y* being a random variable with *m* possible pronunciations y_i with probabilities $p(y_i)$ for grapheme *x*, then the entropy, i.e. the uncertainty about which pronunciation y_i will correspond with *x* is:

$$(1)\ H(Y) = -\sum_{i=1}^{m} p(y_i)\log_2 p(y_i)$$

Given a phoneme *y*, and given variable *X* being a random variable with *m* possible spellings x_i with probabilities $p(x_i)$ for grapheme *y*, then the entropy, i.e. the uncertainty about which spelling x_i will correspond with *y* is:

$$(2)\ H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

An entropy of 0 represents a fully predictable correspondence. The larger the entropy, the less predictable the correspondences.

The average grapheme-to-phoneme entropy for a language *L* with a grapheme inventory consisting of *v* different graphemes is calculated as the average of *v* entropy values of the *v* individual graphemes. Similarly, we calculate the average phoneme-to-grapheme entropy for a language *L* with a phoneme inventory of *w* different phonemes as the average of *w* entropy values corresponding with the *w* individual phonemes.

In order to measure the average grapheme-to-phoneme entropy and the average phoneme-to-grapheme entropy for each language, an R script was developed (R Core team 2016).

3. Corpora and alignment

Due to practical reasons, we decided to include only the 16 Germanic, Romance and Slavic languages that are official EU-languages in our study. The languages in question are: Bulgarian, Croatian, Czech, Danish, Dutch, English, French, German, Italian, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, and Swedish. Unfortunately, this meant excluding one of the languages that is said to have one of the most transparent orthographies, namely Finnish.

3.1 Graphemic corpus

The graphemic corpus, providing us with the official spelling in all 16 languages, consisted of 100 words per language. The words were among the 109 most frequent nouns from the British National Corpus (BNC; Leech, Rayson & Wilson 2001). We excluded words for semantically too similar concepts for our goal, such as the nouns *sort* and *kind*. The remaining 100 nouns were translated into 16 official EU-languages. To make sure that the translators had the same concept in mind, the English source words were presented to them embedded in a sentence. For example, the English noun *state* could be translated into Dutch as ‘toestand’, ‘land’, ‘staat’, or ‘deelstaat’. The provided context was supposed to reduce the ambiguity of the concept behind the source word. All translators were native speakers of the target language. Every list of 100 translations was checked by a second native speaker and, if necessary, corrected. In a few cases of doubt, a third native speaker was consulted.

3.2 Phonemic corpus

The phonemic corpus, providing us with the most standard pronunciation in all 16 languages, consisted of the very same 100 words per language. For every language, the word list was transcribed phonemically using X-SAMPA (Wells 1995) using pronunciation dictionaries.

3.3 Alignments

In a first step, the graphemic and the phonemic transcriptions were aligned word by word in two different ways: (1) As phoneme-to-grapheme alignment, and (2) as grapheme-to-phoneme alignment. These two different alignment tables for all 16 languages serve as the input data for the two entropy measurements, i.e. writing entropy (1) and reading entropy (2).

For (1), three rules were formulated: Rule (i) demanded that every grapheme ought to be put in a separate cell (modelling a reader who is unaware of potential rules for grapheme clusters such as for English <ph>). Rule (ii) stated that phonemes should be aligned in such a way that vowel graphemes represent vowel phonemes and consonant graphemes represent consonant phonemes (modelling the usage of knowledge of the names of the letters by a reader).

Combined, these two rules meant that some graphemes could not be aligned to a phoneme (e.g. the word-final <e> in the English word <time> which is aligned to /taIm/). We presume that the reader who is completely unaware of any phonological rules of the rime <ime> being pronounced VVC rather than VCV is confused by this unpronounced final letter and might try to map it to the preceding consonant.

Therefore, rule (iii) was formulated: No empty cells were allowed in either of the columns, but in cases where no specific phoneme could be aligned to a vowel grapheme, the preceding or following vowel was extended, and when no specific phoneme could be aligned to a consonant grapheme, the preceding or following consonant was ‘prolonged’. This decision aligns ‘silent’ letters to many different phonemes, and thereby results in a higher entropy value for languages with many silent letters. Although we think that this captures the opaqueness of such orthography well, this is arguably not the only way to model the opaqueness of silent letters. Similarly to the problem of unalignable phonemes, there were instances when two phonemes had to be aligned to one single grapheme (recall rule (i) that every grapheme had to be put in a separate cell in the phoneme-to-grapheme alignments). An example of this is the letter <i> in the English word <time> which is aligned to /aI/.

In a very similar way, the phoneme-to-grapheme alignments were conducted. Again, rule (i) required that every phoneme was put in a separate cell, rule (ii) demanded that vowel phonemes were aligned to vowel graphemes and consonant phonemes to consonant graphemes, and rule (iii) stated that empty cells were allowed in none of the two columns. Table 1 shows an example of the alignments for English.

Table 1: Grapheme-to-phoneme alignments (left), which formed the basis of the reading entropy, and phoneme-to-grapheme alignments (right), which formed the basis of the writing entropy in English.

Reading uncertainty		**Writing uncertainty**	
grapheme	*phoneme*	*phoneme*	*grapheme*
t	t	t	t
i	aɪ	a	i
m	m	ɪ	i
e	m	m	me

4. Results

The results from the entropy measurements are given in Figure 1. It becomes obvious that English is the language with the least predictable orthography from a beginners’ point of view, both in writing and reading. French has a very opaque orthography for a beginning reader, and it is still rather hard to spell properly if nothing more is known than the names of the letters. Regarding the uncertainty of spelling, German and Danish share a third place with fairly opaque phoneme-to-grapheme representations, and Danish grapheme-to-phoneme representations are even more opaque than the French. On the other end of the continuum the differences in transparency are less pronounced. Apart from Swedish, only Romance and Slavic languages have entropy values of less than 0.5. Among the 16 included languages, Bulgarian is the language that is easiest to read and Croatian has the orthography which is easiest to write.

5. Conclusion

Using an R-script that calculates the entropy (uncertainty) of grapheme-to-phoneme and phoneme-to-grapheme correspondences, we modelled a beginning reader and writer, who is familiar with nothing more than the letter names in every orthography. For this type of illiterate learner, English and French are by far the most opaque orthographies to read, and English and Danish are the most opaque orthographies to write, while most Slavic and Romance languages are far less opaque.

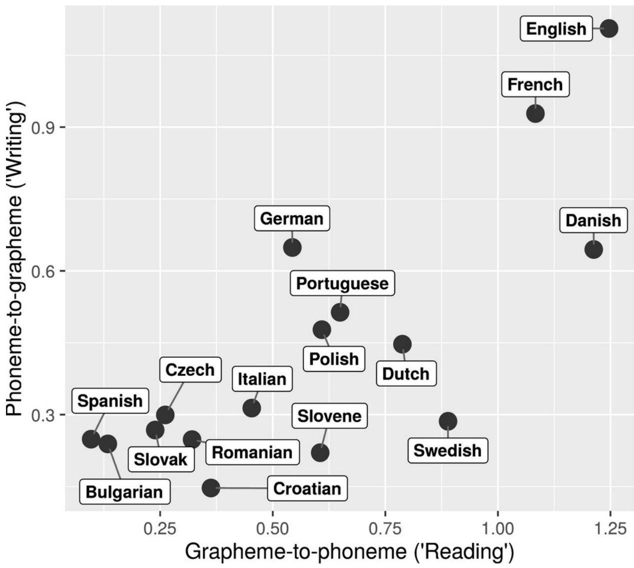

Figure 1: Grapheme-to-phoneme (‘reading’) entropy values plotted against phoneme-to-grapheme (‘writing’) entropy values for all 16 languages.

The main reason why Slavic languages as a group are characterized by relatively predictable grapheme-to-phoneme and phoneme-to-grapheme correspondences are the orthographic reforms mostly carried out in the 19th century. Czech was the first language to standardize its orthographic system and the orthographies of Slovak, Slovene and Serbo-Croatian used it as a model when creating their own orthographies. The rule attributed to German philologist Adelung “write as you speak and read as it is written” was often explicitly invoked by Slavic language experts carrying out the reforms. Therefore, the current orthographic consistency in Czech, Slovak, Slovene and Croatian is a result of a conscious effort to ensure transparency. Bulgarian uses the Cyrillic alphabet and while the written language looks completely different compared to the other 15 languages we studied, it boasts the lowest entropy values. One notable exception in terms of transparency of Slavic languages is Polish. Ever since the Latin alphabet was adopted in Polish, it was clear that it could not accommodate palatal and retroflex consonants as well as the nasal vowels, typical of this language. Subsequent reforms did improve things to an extent, but due to numerous complex and often inconsistent rules, Polish orthography remains the least transparent one among the Slavic languages.

Interesting is also the very different degree of transparency of the Scandinavian orthographies that were included in our study, i.e. the orthographies of the two very closely related languages Danish and Swedish. The extremely differing entropy values can be explained by two processes, as summarised by Elbro (2005: 33): Firstly, “Danish orthography was already old when a national norm was first established around the year 1200”, and secondly, “things have become worse since the 1200s” as “spoken Danish has changed more than most Germanic languages since the 1200s”. In other words, while the Danish spelling norm always has preserved a rather archaic pronunciation, the spoken language has developed even faster than in many other langusages and the spelling has not been adjusted to these changes. In Danish, many sounds that have been lost in pronunciation are still preserved as silent letters, such as the spelling <mild> for /milˀ/, or the spelling of <tolv> for /tɔlˀ/, or the spelling <lærere> for /lɛːɔɔ/. In comparison, these words are spelled <mild>, <tolv> and <lärare> in Swedish, and pronounced /mɪld/, /tɔlv/ och /lærarə/. Also loan words have different appearances in the two languages: Danish has more foreign words than Swedish, as many loan words into Swedish are translated (rendering *calques*). An example for this is the Danish word *weekend* versus the Swedish calque *veckoslut* or the Old Norse word *helg* (from *helig*, Engl. ‘holy’). Furthermore, even for directly loaned words, Danish has a tendency to preserve the foreign spelling such as in Danish <niveau> versus Swedish <nivå> or Danish <restaurant> versus Swedish <restaurang>).

Let us now return to the study conducted by Seymour et al. (2003), who measured literacy acquisition in 13 orthographies and interpreted their results on the basis of a "hypothetical classification of participating languages" (Seymour et al. 2003: 146). They classified the 13 orthographies on a five-point scale from *shallow* to *deep*. Importantly, they did not make a difference between grapheme-to-phoneme depth and phoneme-to-grapheme depth, as we did. Recalling that they elicited data from reading development only, we might assume that their classification is meant mainly or exclusively for *reading* opaqueness, and should be compared to our grapheme-to-phoneme measurements in the first place. Although their classification is very rough (with e.g. Italian, Spanish and German in one (nameless) category representing *semi-shallow*, and Portuguese, Dutch and Swedish in one (equally nameless) category between shallow and deep, their ranking is completely supported by our results.

This study represents a commensurate comparison of phoneme-to-grapheme correspondence and vice versa in 16 European languages. The findings can be taken as a predictor of reading and spelling difficulties in each of the languages, but can also serve as background information for psycholinguistics experiments. The fact that we modelled a beginning reader and writer with hardly any orthographical knowledge limits the validity of this study. We opted for this setup since it is vital for a cross-linguistic study to take the same criteria as a basis, and the criteria we used were easily applicable to the 16 languages. Modelling a slightly more advanced reader and writer, as Borgwaldt, Hellwig & De Groot (2005) did for Dutch, English, French, German, Hungarian, Italian and Portuguese, is a very useful additional approach. Another useful extension of the present study would be the inclusion of more languages.

References

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

Chomsky, N. & Halle, M. (1968). *The sound patterns of English*. New York: Harper & Row.

Elbro, C. (2005). Literacy acquisition in Danish: A deep orthography in cross-linguistic light. In: R. Malatesha Joshi & P. G. Aaron (eds), *Handbook of orthography and literacy*, 31-45.

Frith, U., Wimmer, H., & Landerl, K. (1998). Differences in phonological recoding in German- and English-speaking children. *Journal of the Society for the Scientic Study of Reading*, 2, 31–54.

Goswami, U., Porpodas, C., & Wheelwright, S. (1997). Children's orthographic representations in English and Greek. *European Journal of Psychology of Education*, 12, 273–292.

Goswami, U., Gombert, J. E., & de Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics*, 19, 19–52.

Goswami, U. (2008). The development of reading across languages. Annals of the New York Academy of Sciences, 1145, 1–12.

Katz, L., & Frost, R. (1992). The reading process is different for different orthographies: the orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning: Advances in psychology*, 94, 67–84. Oxford: North-Holland.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

Nicolai, G. & Kondrak, G. (2015). English Spelling is not "close to optimal". Paper presented at the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015), Denver, CO.

Peereman, R., & Content, A. (1996). Orthographic and phonological neighbourhoods in naming: Not all neighbors are as mighty in orthographic space. *Journal of Memory and Language*, 37, 382–410.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

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

Shannon, Claude E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*. 27 (3): 379–423.

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.

Wells, J. C. (1995) *Computer-coding the IPA: a proposed extension of SAMPA*. UCL Phonetics and Linguistics.

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition*, 51, 91–103.

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: a cross-language investigation. *Psychological Science*, 21, 551–559.

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

Ziegler, J. C. & Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin*, 131, 3–29.

Ziegler, J. C. & Goswami, U. (2006). Becoming literate in different languages: similar problems, different solutions. *Developmental Science*, 9, 429–453.