Unpredictability and complexity of print-to-speech correspondences increase reliance on lexical processes: More evidence for the Orthographic Depth Hypothesis

Xenia Schmalz[1,2], Elisabeth Beyersmann[3], Eddy Cavalli[3], & Eva Marinus[2,4]

[1] DPSS, Università degli Studi di Padova, Italy

[2] ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, Australia

[3] Laboratoire de Psychologie Cognitive, Aix-Marseille Université, France

[4] Department of Cognitive Science, Macquarie University, Australia

Abstract

The Orthographic Depth Hypothesis (Katz & Frost, 1992) proposes cross-linguistic differences in the involvement of lexical processing during reading. In orthographies with complex, inconsistent, and/or incomplete sublexical correspondences, decoding is more difficult and therefore slower. This gives more time to the lexical route to retrieve information, and leads to a greater ratio of lexical processing. We test whether this mechanism applies both for words with inconsistent (in English) and for words with complex (in French) correspondences. As complex correspondences are sufficient to derive a correct pronunciation, an increase in lexical processing may not occur. In a reading aloud task, we used the frequency effect to measure lexical processing. The data showed stronger involvement of lexical processing for inconsistent compared to consistent words, and for complex compared to simple words. The results confirm that Katz and Frost's proposed mechanism applies to different sources of orthographic depth.

Keywords:

Dual-route model, cross-linguistic, French, English.

Unpredictability and complexity of print-to-speech correspondences increase reliance on lexical processes: More evidence for the Orthographic Depth Hypothesis

Previous research has shown that the cognitive processes underlying reading differ across orthographies. This is true for both adult reading (Frost, Katz, & Bentin, 1987; Rau, Moll, Snowling, & Landerl, 2015; Schmalz et al., 2014) and reading development (Aro & Wimmer, 2003; Landerl, Wimmer, & Frith, 1997; Mann & Wimmer, 2002; Marinus, Nation, & de Jong, 2015; Seymour, Aro, & Erskine, 2003). From a theoretical perspective, it is important to isolate how and why cognitive mechanisms of reading differ across orthographies. This will provide insight into how the universal perceptual systems interact with specific properties of each language and orthography, and lay out benchmarks for models of reading (Frost, 2012; Share, 2008).

From the onset of research on cross-linguistic differences in reading, the concept that has received the most attention is orthographic depth (Bridgeman, 1987; Frost et al., 1987; Katz & Feldman, 1983; Turvey, Feldman, & Lukatela, 1984). Broadly speaking, orthographic depth refers to the cross-linguistic variability in the closeness of the relationship between orthographic word forms and their pronunciations. Within most models of reading, print-to-speech correspondences are important for a process called sublexical decoding, whereby the pronunciation of a word or nonword is assembled using the knowledge of the regularities that underlie print-to-speech conversion in a given orthography (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Perry, Ziegler, & Zorzi, 2007; Plaut, McClelland, Seidenberg, & Patterson, 1996). In shallow orthographies, such as Italian, this is a relatively simple process. The Italian grapheme *a*, for example, almost always maps onto the phoneme

/a/. Conversely, in English, a prototypical deep orthography, the grapheme *a* can be pronounced as in “cat”, “nation”, “wasp”, or “false”. Part of this ambiguity can be resolved by developing sensitivity to more complex regularities that exist in the orthography. For example, the pronunciation of the word *wasp* is predictable, because an *a* preceded by a *w* tends to be pronounced as /ɔ/ (Schmalz et al., 2014; Treiman, Kessler, & Bick, 2003). In English orthography, however, there are also instances where the pronunciation is not predictable based on any sublexical information, such as in the words *yacht* or *colonel*. Therefore, there are two reasons why English orthography might be considered deep: first, the relatively high degree of complexity of the print-speech correspondences compared to orthographies such as German, Italian, and Dutch, and second, a high degree of unpredictability, even when those complex rules are applied (Schmalz, Marinus, Coltheart, & Castles, 2015; van den Bosch, Content, Daelemans, & de Gelder, 1994; Ziegler, Perry, & Coltheart, 2000).

Corpus analyses have shown that, across orthographies, unpredictability and complexity are dissociable on a linguistic level (Schmalz et al., 2015; van den Bosch et al., 1994): although orthographies with simple correspondences tend to also have a high degree of predictability, these two concepts are not perfectly correlated. A particularly interesting example is French orthography, as there is a discrepancy between the degree of complexity and unpredictability. Specifically, French is high in complexity, because it contains multi-letter rules (*au* → /o/) and context-sensitive rules (*c[a,o,u]* → /k/, *c[e,i]* → /s/), but low in unpredictability (Schmalz et al., 2015; van den Bosch et al., 1994).

To date, it is unclear whether complexity and unpredictability of the sublexical correspondences act as separate sources of orthographic depth, or if they affect reading processes in the same way on a behavioural level. An existing hypothesis on

orthographic depth is the Orthographic Depth Hypothesis (hereafter: ODH; Katz & Frost, 1992). Here, the authors offer both a well-specified definition of orthographic depth, and propose a specific cognitive mechanism that drives cross-linguistic differences as a function of orthographic depth. In deep orthographies, they describe print-to-speech conversion code as characterised by *complex*, *inconsistent*, and/or *incomplete* sublexical information. This makes the sublexical conversion process more difficult in deep compared to shallow orthographies. As a result, the sublexical conversion process is impaired in one way or another, which gives more time for a lexical look-up mechanism to derive the correct pronunciation. This leads to a higher overall ratio of lexical-to-sublexical processing, as a function of degree of orthographic depth.

It is particularly noteworthy that Katz and Frost (1992) list three different properties that underlie the sublexical regularities of deep versus shallow orthographies: complexity, consistency, and incompleteness. The concepts of complexity and consistency map onto the distinction between complexity and unpredictability proposed by Schmalz et al. (2015; see also van den Bosch et al., 1994). Yet despite the theoretical and linguistic work that has shown a distinction between these multiple constructs underlying orthographic depth, whether these may differentially affect reading processes on the behavioural level has not been previously empirically tested.

The first construct proposed by both Katz and Frost (1992) and Schmalz et al. (2015) is complexity. An orthography with complex correspondences is characterised by multi-letter rules, where several letters are required to denote a single phoneme (e.g., *augh* → /o:/ in English, *aient* → /ɛ/ in French) and/or context-sensitive regularities, where surrounding letters affect a grapheme's pronunciation (e.g., in

English, *a* is pronounced as /ɔ/ when preceded by a *w*, as in "swan"; in French, a *g* is pronounced as /ʒ/ when followed by an *i* or *e,* as in *gélatine*). When words contain complex correspondences, the sublexical information is sufficient to access full information about the word's phonology and semantics, once these complex rules are applied. However, evidence exists that applying multi-letter rules slows down the sublexical procedure, as they cause a conflict between the pronunciation of the single letters (e.g., in English, *t* and *h*) and the grapheme's pronunciation, *th* → /θ/ (Marinus & de Jong, 2010; Rastle & Coltheart, 1998; Rey, Jacobs, Schmidt-Weigand, & Ziegler, 1998).

The second construct that has been described by both Katz and Frost (1992) and Schmalz et al. (2015) is inconsistency, or unpredictability. Inconsistency is the presence of two or more pronunciations for the same orthographic unit. Conventionally, this is defined at the level of a word's body (e.g., the body *–ear* is inconsistent because it can be pronounced as in "hear" or "bear"), but the same measure can also be applied to graphemes (e.g., the grapheme *th* is inconsistent, because it can be pronounced as in "thistle", "this", or "thyme"). For this source of depth, the sublexical information is not sufficient to derive the correct pronunciation. For example, the English words "tough", "though" and "through" have nearly identical sublexical information, but each of them has a different pronunciation of the grapheme *ough*, which cannot be derived without knowledge of the whole word (see Schmalz et al., 2015, for an indepth discussion). According to rule-based computational models, such as the Dual Route Cascaded (DRC) model (Coltheart et al., 2001), such words need to be read aloud via the lexical route for a correct response, because the sublexical route will give a "regularised", or rule-based response for words which do not comply to the rules (e.g., /θaim/ for the word

"thyme"). In a connectionist framework (Perry et al., 2007; Plaut et al., 1996), such words require the reliance on larger units, which in the case of unpredictable words coincide with a whole word (e.g., *ough* is pronounced as in "tough" when preceded by a *t* and as in "though" when preceded by a *th*). Arguably, the fact that the orthographic unit coincides with a whole word makes its processing qualitatively different from processing sublexical orthographic units, as whole words have direct connections to their semantic information, while sublexical units do not (for a discussion, see Schmalz et al., 2015).

The third construct proposed by Katz and Frost (1992) is incompleteness. This construct is of high relevance to Semitic orthographies, where vowels are not always represented. In pointed Hebrew, the sublexical information is complete, because all phonemes are represented; vowels are represented as diacritics. Generally, however, texts are written in unpointed Hebrew, without vowel markings. Here, vowel information is incomplete, and the pronunciation needs to be derived via semantic context by fluent readers. Incompleteness is not of high relevance for European orthographies, however. In the European alphabetic scripts, the orthographic (sublexical and whole-word) information is mostly sufficient to assemble a full phonological representation and to use this to access a word's semantics. There are some examples of words with incomplete lexical and sublexical information, namely heterophonic homographs. For a word like "present", semantic context is needed to derive both a pronunciation, and to access different semantic information depending on whether this word occurs as a verb or a noun. By definition, lexical-semantic processing is required when the sublexical correspondences are incomplete.

According to the ODH, complexity, inconsistency, and incompleteness result in a higher ratio of lexical and/or semantic to sublexical processing. The notion of an

independent lexical and sublexical route is the basis of the dual-route framework (Coltheart et al., 2001; Perry et al., 2007). Here, the lexical and sublexical routes operate in parallel to obtain a pronunciation from an orthographic input. The longer the sublexical route takes, the more the final pronunciation will be influenced by excitatory connections from the orthographic lexicon to the phonological lexicon and to the phonological output buffer. If the sublexical information can be processed quickly, the phonological output will be driven to a greater extent by phoneme activation from the sublexical units.

Previous research has provided support for a stronger lexical influence for deep compared to shallow scripts, as predicted by the ODH. Frost et al. (1987) showed, in a between-language comparison, that lexical and semantic marker effects increase as a function of depth in Serbo-Croatian (a shallow orthography), English (medium) and unpointed Hebrew (deep). In a further study, Frost (1994), took advantage of the presence of both the shallow pointed and the deep unpointed script in Hebrew. This allows for a within-item design, where the same words can be presented with and without diacritics. Again, Frost (1994) showed stronger lexical (word frequency) and semantic (semantic priming) effects for the deep compared to the shallow script.

Both studies support the view that incompleteness increases the reliance on lexical processing, as both report a comparison of unpointed Hebrew with a complete orthography (pointed Hebrew, and English and Serbo-Croatian). The comparison between English and Serbo-Croatian, however, can be interpreted in different ways, because these two orthographies differ from each other both in terms of complexity and unpredictability. The first possibility, which is in line with the ODH, is that complex correspondences slow down the process of sublexical decoding. Thus, while

the sublexical output is in principle sufficient for a correct response to occur, the slow-down will allow more time for the lexical route to contribute to the final phonological output. This would mean that any source of orthographic depth (i.e., complexity, unpredictability, or incompleteness) should increase the relative contribution of the lexical route.

Alternatively, it is possible that there is a qualitatively different impact of unpredictability and incompleteness as compared to complexity: As unpredictability and incompleteness make it impossible for the reader to compute a pronunciation from the sublexical information, the final response of the sublexical route will be either incorrect or partial. In this case, a correct reading aloud response cannot occur until the lexical route has provided enough activation to the phonological output buffer. This is different for words with complex correspondences: here, the sublexical information is, in principle, sufficient for a correct pronunciation. Any slow-down associated with the presence of complex correspondences might not be sufficient to result in a substantial effect on the relative amount of lexical processing.

The existing studies do not allow us to differentiate between the two possibilities. To our knowledge, all comparisons of lexical/semantic marker effects used orthography pairs which differ both in terms of complexity and unpredictability, such as English and Serbo-Croatian (Frost et al., 1987) or English and German (Frith, Wimmer, & Landerl, 1998; Rau et al., 2015). The main aim of the current study was to distinguish between these two possibilities. We use two orthographies, where the correspondences reflect two different sources of depth, namely unpredictability in English, and complexity in French (Schmalz et al., 2015; van den Bosch et al., 1994). We chose reading aloud rather than silent reading as the experimental task, because the ODH is specifically concerned with the process of deriving speech from print.

Lexical decision is considered to be less sensitive to this sublexical process, as high accuracy on this task can be achieved purely by relying on lexical access (Coltheart et al., 2001).

The Unpredictability Measure

Defining unpredictability is not straightforward, because existing models of reading make different assumptions about the way in which the sublexical route assembles a pronunciation (Coltheart et al., 2001; Plaut et al., 1996; for a discussion, see Schmalz et al., 2015). Given that there is no consensus about the type of information that is used to assemble a pronunciation, it is also unclear what kinds of words would be considered to have an unpredictable pronunciation. To ensure that the results are meaningful beyond the assumptions of a specific model, we use a definition which is compatible with both connectionist and rule-based models: we classify a word as unpredictable, if (1) it is both irregular (by the set of grapheme-phoneme correspondence rules implemented within the rule-based computational model, DRC; Coltheart et al., 2001) and inconsistent (i.e., if the body has more than one possible pronunciation), such as the word "ghost", or (2) if it is irregular, and does not have any body neighbours, such as the word "debt". Thus, neither grapheme-phoneme correspondences nor body-rime correspondences can be reliably used to read aloud these words correctly.

The concepts of irregularity and inconsistency are strongly correlated, but reflect theoretically different constructs and can be manipulated to vary orthogonally (Andrews, 1982; Cortese & Simpson, 2000; Jared, 1997, 2002; Jared, McRae, & Seidenberg, 1990). Here, we classified words as predictable if the pronunciation was predictable both from grapheme-phoneme correspondence rules (i.e., regular) and from body-rime correspondences (i.e., consistent), and as unpredictable when neither

source could be used to read aloud the words correctly. We excluded words that are regular but inconsistent (e.g., "mint", which is regular but has the enemy "pint") or irregular but consistent (e.g., "walk", which should be pronounced as /wælk/ according to the DRC). Recent behavioural data suggests that participants rely on information from various types of sources to predict a novel word's pronunciation (Schmalz et al., 2014); as it is not yet clear how the cognitive system merges conflicting information from different sources, we excluded these types of words for the current purposes.

It is unconventional to use predictability as a variable in psycholinguistic research. To date, the literature has focussed predominantly on contrasting the effects of regularity with those of consistency (Andrews, 1982; Cortese & Simpson, 2000; Jared, 1997, 2002; Jared et al., 1990). Rule-based models, such as DRC, predict effects of regularity, because a lack of compliance to grapheme-phoneme correspondence rules should impair the reading aloud process via the sublexical route. Connectionist models, such as the triangle or connectionist dual processing (CDP) models (Perry et al., 2007; Plaut et al., 1996), use a learning algorithm to extract the relationships between print and speech, which becomes more difficult when a given orthographic pattern can map onto multiple pronunciations (e.g., the body *–ost*, which can be pronounced as in "ghost" or as in "lost"). Thus, connectionist models predict an effect of consistency, but not regularity. While previous studies have shown that inconsistent words and nonwords are read aloud more slowly than matched consistent items (Andrews, 1982; Cortese & Simpson, 2000; Glushko, 1979; Jared, 1997, 2002; Jared et al., 1990), other data suggests that participants also rely on print-to-speech rules, especially for unusual orthographic patterns (Andrews & Scarratt, 1998; Pritchard, Coltheart, Palethorpe, & Castles, 2012; Robidoux & Pritchard, 2014).

The current study

In Experiment 1, we compare the frequency effect for English words with predictable versus unpredictable correspondences. The English orthography is used, because the relatively high degree of both complexity and unpredictability allows us to manipulate frequency and predictability. If there is stronger involvement of lexical processing when the pronunciation of a word is unpredictable, we expect a frequency-by-predictability interaction, where the frequency effect is larger for unpredictable compared to predictable words. This study serves as a conceptual replication of the finding of Frost et al. (1987) that there is a stronger relative involvement of the lexical route when the correspondences are unpredictable compared to when they are predictable.

In Experiment 2, we use the French orthography, which has a high degree of complexity, while being highly predictable. This allows us to manipulate frequency and complexity in a within-subject design and without unpredictability as a confounding variable. We aim to establish whether there is a stronger frequency effect for words containing complex correspondences, compared to words which contain only simple correspondences (i.e., there is a one-to-one correspondence between letters and sounds). If the complexity of the correspondences slows down the assembly process, we expect to find a frequency-by-complexity interaction, as lexical processing should be stronger for words with complex than simple correspondences according to the ODH. If we obtain this pattern, it would indicate that both complexity and unpredictability affect reading processes in adults in the same way as incompleteness in the previous studies (Frost, 1994; Frost et al., 1987). If we do not find a frequency-by-complexity interaction, this means that cross-linguistic

differences in the relative reliance on lexical processing are driven by unpredictability and incompleteness, but not complexity.

Experiment 1: Unpredictability in English

Method

Participants

Twenty undergraduate students at an Australian university participated in the experiment. All were native speakers of English and received course credit for their participation.

Items

To justify the use of the predictability metric rather than the more conventional consistency metric, we first verified that predictability reflects a psychologically valid construct. We created two models from the full dataset that is analysed in Experiment 1 (see below). The models were nearly identical to those described in the Results section below. The independent variables were predictability (coded as a binary contrast) or consistency (centred ratio of friends to enemies), centred log frequency, and the two-way interaction. Note that we centred all continuous independent variables (by subtracting each value from the mean) and contrast-coded dichotomous conditions (as 0.5 and -0.5) because LME provides parameter estimates as deviations from the point closest to zero rather than deviations from the mean. The dependent variable was trimmed inverse RT (for more details about the trimming procedure, see the Results section below). Items and participants were included as random effects, and the slope of the frequency effect was allowed to vary across participants (Barr, Levy, Scheepers, & Tily, 2013). The model with predictability as the independent variable yielded a numerically better fit than the

model with the consistency ratio as the independent variable (AIC = 4620 for the former, AIC = 4658 for the latter). A Bayesian analysis, where the two models were contrasted, provided support for the model which used predictability as the independent variable over the model using consistency, with a Bayes Factor value > 1,000,000 (for a description of how we interpret Bayes Factors, see below). This justifies the use of predictability as an independent variable, and suggests that predictability has stronger psychological validity than consistency.

For the experiment, we used only monosyllabic words, because traditionally, measures of regularity and consistency, which form the basis of the predictability construct, have been defined for monosyllabic words only (but see Chateau & Jared, 2003; Kearns et al., 2014; Yap & Balota, 2009, for an extension of the consistency measure to multisyllabic words). We extracted all monosyllabic words from the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012). We retained all words with log frequencies between 0 and 2, because analyses of large-scale lexical databases have shown that the frequency effect is most robust for this log frequency range (Balota et al., 2007; Brysbaert et al., 2011; Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010). All words had a lexical decision accuracy > 80%, suggesting that the words should be familiar to the majority of undergraduate students. We classified the words as predictable or unpredictable based on the above criteria.

We selected a total of 376 words. Half of these were predictable (e.g., “forge”) and half were unpredictable (e.g., “ghost”), and they were chosen to vary in frequency, as half had a relatively low frequency (log frequency of 0 – 1) and the other half a relatively high frequency (log frequency of 1 – 2). Note that we treat frequency as a continuum rather than a dichotomy throughout the paper to increase

experimental power. Frequency, as well as orthographic N counts, are based on the subtitle counts provided by the BLP (Keuleers et al., 2012; van Heuven, Mandera, Keuleers, & Brysbaert, 2014). Linear models were performed to assess whether any of the item characteristics co-varied with frequency or unpredictability. In separate analyses, each centred potential covariate was used as the dependent variable; centred frequency, contrast-coded predictability, and their interaction were used as predictor variables. The outcomes of this set of analyses are shown in Table 1. The individual items and their full descriptives, as well as the raw data and the R script used for the current study, can be found here: osf.io/hm8fw. Note that orthographic neighbourhood and Phonological Levenshtein Distance (the number of phoneme substitutions, deletions, or additions which are required to reach the nearest 20 neighbours; see Yarkoni, Balota, & Yap, 2008) co-vary with frequency, and the ratio of letters to phonemes differs across predictable and unpredictable words. However, the critical comparison in the current experiment is the interaction between predictability and frequency, and none of the covariates show a stronger manipulation for the predictable than unpredictable condition, all $p > 0.3$ for the interaction. We therefore do not include any of them as covariates in the main analysis. To confirm that these potential confounds do not influence the results, however, we present a covariate analysis in a post-hoc test.

TABLE 1 ABOUT HERE

Procedure

Item presentation was controlled with DMDX (Forster & Forster, 2003). The words were shown, one at a time, in random order, for 2.5 seconds or until the voice-key was triggered. The participants were instructed to read aloud each item as quickly and accurately as possible.

Results and discussion

The reading aloud responses were scored offline with the software CheckVocal (Protopapas, 2007), as correct, incorrect, or no response. Response latencies were readjusted using CheckVocal, based on the onset of the sound waves, in the case of premature of late voice key triggers. This removes potential biases associated with first phonemes.

The data were further analysed using the software R, both with Linear Mixed Effect (LME) models (Baayen, Davidson, & Bates, 2008) and with Bayes Factors (Morey & Rouder, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009). LMEs allow us to obtain an estimate of the slope (which serve as descriptives given the use of a continuous measure of frequency). We provide the results of *t*-tests, and *p*-values, when appropriate, to provide a point of reference for those unfamiliar with Bayesian analyses.

We report Bayes Factors for all theoretically interesting comparisons (i.e., for the critical interactions), and base our conclusions on them. Unlike frequentist statistics, Bayes Factors allow us to quantify the evidence for (or against) an effect or interaction of interest, given a prior belief. Therefore, they arguably provide a closer link to the conclusions that can be drawn from the data. Here, we use the default prior of the BayesFactor package, which assumes a Cauchy distribution around the effect size of $r = 0.5$ (Morey & Rouder, 2014). We interpret the results according to a set of guidelines described in Rouder et al. (2009): Bayes Factor values smaller than 1/3 provide evidence against an effect or interaction, values between 1/3 and 1 and between 1 and 3 are considered to provide anecdotal or equivocal evidence against or for it, respectively, values larger than 3 provide some evidence for the effect or interaction, and values larger than 10 provide strong evidence. Thus, throughout the

paper, smaller values provide evidence for a null hypothesis, and larger values provide evidence for the alternative hypothesis.

For the LME model, we used inverse RTs as the dependent variables. For the independent variables, we used centred log frequency as a continuous predictor, and predictability, contrast-coded as 0.5 (predictable) and -0.5 (unpredictable), as a binary predictor. The model also included previous RT (Baayen, 2008). Participants and items were included as random factors, and the frequency slope was allowed to vary across participants (Barr et al., 2013).

There were seven non-responses (0.01% of all data), and overall accuracy was 97.2%. The accuracy rates ranged from 93.4% for low-frequency unpredictable words to 99.3% for low-frequency predictable words. An LME model with accuracy as the dependent variable showed a main effect of predictability, $\beta = 1.5$, $z = 3.9$, $p < 0.0001$, reflecting higher accuracy for the predictable (99.1%) than unpredictable (95.3%) conditions[1]. The interaction between frequency and predictability was significant, $\beta = -1.7$, $z = -2.5$, $p = 0.013$, indicating a facilitatory frequency slope for unpredictable ($\beta = 1.5$) but not predictable ($\beta = -0.3$) words. The main effect of frequency was not significant, $p = 0.1$. The results are broadly in line with the RT results discussed below. As the error rate was relatively high, and there was not a lot of variability between the conditions, we draw conclusions based on the RT results only.

Before conducting the RT analyses, we excluded all incorrect responses, and trials with latencies < 300 ms (0.2% of the data) and > 1200 ms (0.1% of the data). This yielded an approximately normal distribution of inverse RTs. When we

[1] In the accuracy analyses, we did not allow slope to vary across participants, because this model failed to converge. This is a common issue with maximal models (see Bates, Kliegl, Vasishth & Baayen, 2015). However, as the accuracy data were not interpreted, this is not an issue for the current study.

artificially dichotomise frequency into high (log frequency > 1) and low (log frequency < 1), the averages RTs of the trimmed dataset are 494.2 (SD = 97.8) and 494.1 (SD = 97.2), respectively, for the predictable words, and 503.6 (SD = 97.5) and 528.3 (SD = 115.1), respectively, for the unpredictable words.

The LME showed a significant main effect of predictability, β = -0.08, t = -5.3, p < 0.0001, reflecting shorter RTs for predictable ("forge") than unpredictable ("ghost") words, and a main effect of frequency, β = -0.05, t = -3.5, p = 0.0006, indicating shorter RTs for words of higher frequencies. The interaction was also significant, β = 0.07, t = 2.3, p = 0.02, with a steeper frequency slope for unpredictable compared to predictable words. In a comparison of the full model against an additive one that included the main effects of predictability and frequency, the Bayes Factor provided anecdotal evidence for the presence of the interaction, BF = 1.7 (±1.2%).

Follow-up analyses

To potentially strengthen the case for the interaction, we retrieved all trial-level data for our items from the English Lexicon Project reading aloud database (Balota et al., 2007). Note that we included data from the ELP and not the BLP because the ELP has both lexical decision and reading aloud data, while the BLP only has lexical decision. As both the ELP and our experiment employed a standardised reading-aloud procedure, we can increase the amount of evidence by collapsing the two datasets.

The ELP contains trial-level reading aloud data for 375 of the original 376 words. These include 10342 valid and correct trials, with an average of 27.6 participants per word. We combined the data from our experiment with data of the ELP. The trimming procedure of this bigger item set was identical to that of the

original data, as was the model, except that previous RT was not included, as it was unavailable in the ELP database. Dichotomising frequency, the average RTs for the four types of items were 561.9 ms (SD = 125.4) for high-frequency predictable words, 570.2 (SD = 125.7) for high-frequency unpredictable; 575.3 (SD = 132.4) for low-frequency predictable, and 597.6 (SD = 141.5) for low-frequency unpredictable. Note that *p*-values are not reported for any of the follow-up analyses in this paper, as due to the multiple comparisons, the Type-I error rate increases and is no longer 5% with a cut-off of $\alpha = 0.05$ (Cramer et al., 2015; Simmons, Nelson, & Simonsohn, 2011).

The LME results showed the same pattern as the original data, with shorter RTs for more frequent compared to less frequent words, with a slope of $\beta = -0.07$, $t = -10.0$, with shorter latencies for predictable than unpredictable words, $\beta = -0.05$, $t = -6.6$, and a steeper frequency slope for unpredictable than predictable words by $\beta = 0.05$, $t = 3.3$. Importantly, the Bayes Factor now provided evidence for the presence of the interaction between frequency and predictability, BF = 9.6 (±0.6%).

In an additional post-hoc analysis, we ensured that the obtained results remain stable after taking into account potential confounds. As shown in Table 1, some of the psycholinguistic variables co-varied with our manipulations. The model was identical to the one above, but we also included main effects of orthographic N and PLD20 (which differ as a function of frequency), and the ratio of letters to phonemes (which differs as a function of predictability). The adjusted means – when frequency is dichotomised – are 606.5 ms for high-frequency predictable words, 612.6 ms for high-frequency unpredictable words, 624.7 ms for low-frequency predictable words, and 646.5 ms for low-frequency unpredictable words. The results of the full model can be downloaded from the OSF folder (osf.io/hm8fw). The patterns of results did not change: the LME showed a main effect of frequency, $\beta = -0.07$, $t = -10.0$,

predictability, β = -0.04, t = -5.8, and the interaction, β = -0.05, t = 3.8. The Bayes Factor provided evidence for the presence of the critical interaction, BF = 50.7 (±0.88%).

In sum, we found strong evidence for the predicted interaction between frequency and predictability, where the frequency effect is stronger for unpredictable than predictable words. This provides a conceptual replication of previous experiment by Frost and colleagues (Frost, 1992; Frost et al., 1987), and evidence for the ODH. Specifically, the results suggest that unpredictability of print-to-speech correspondences impairs sublexical processing, which results in stronger lexical involvement compared to words with predictable correspondences.

Experiment 2: Complexity in French

For French, the aim was to assess whether the frequency effect for words containing complex print-to-speech correspondences is stronger than for words where the pronunciation can be deciphered based on simple single-letter correspondences. This would provide further support for the ODH, and insights about the orthographic characteristics that may lead to a script being classified as deep or shallow. A lack of an interaction between frequency and complexity would suggest that complex correspondences are processed qualitatively differently to unpredictable correspondences.

Method

Participants and procedure

The participants were 24 students from a university in France. All were native speakers of French and received course credit in exchange for their participation. The procedure was identical to Experiment 1.

Items

We retrieved words and their corresponding information from the Lexique 2 database (New, Pallier, Brysbaert, & Ferrand, 2004) and the French Lexicon Project (Ferrand et al., 2010). For frequency, we relied on subtitle counts (Brysbaert et al., 2011; Brysbaert & New, 2009; New, Brysbaert, Veronis, & Pallier, 2007). We again removed words with log frequencies of < 0 or > 2. To classify words as complex or simple, we used the ratio of letters to phonemes in each word: the presence of multi-letter correspondences means that multiple letters correspond to a single phoneme, thus a complex word has a letter-to-phoneme ratio > 1. Simple words were those with a letter-to-phoneme ratio of one (e.g., "garnir"), and words with a ratio of greater than one were considered complex (e.g., "gâteau"). In the database, this procedure classified 280 words (8.9%) as "simple", and 2852 (91.1%) as "complex".

We selected 384 words, half with complex correspondences ("gâteau") and half with simple correspondences ("garnir"). The words were chosen to vary in frequency, where half the items had frequency counts lower than 1, and the other half higher than 1. In addition, the items were chosen such that they did not differ, across conditions, on average grapheme consistency (Lété, Sprenger-Charolles, & Colé, 2004), suggesting that there were no differences in the degree of unpredictability. Overall, the French orthography has a high degree of predictability once complex rules are taken into account (Schmalz et al., 2015; Ziegler, Perry & Coltheart, 2003). However, there are some words with ambiguous pronunciations (e.g., "femme", where the second letter is pronounced as /a/ rather than the default /ɛ/). While, to our knowledge, there is no quantification method of regularity that can be applied to polysyllabic words in French, the Manulex database contains average grapheme consistency ratings (Lété et al., 2004). We use these as a measure of unpredictability,

as words with unpredictable pronunciations necessarily have graphemes that can be pronounced in multiple ways.

All items had a lexical decision accuracy, according to the FLP, of > 80%. The descriptive statistics are listed in Table 2. For the full item set with individual word characteristics, as well as the raw data and R scripts, see here: osf.io/hm8fw. Again, in the results section, we will follow up with covariate analyses to ensure that the results cannot be explained by the variables that differ as a function of the manipulation.

TABLE 2 ABOUT HERE

Results and discussion

The data were scored with CheckVocal as correct, incorrect, or no response, and the RTs were adjusted when the voice-key had been triggered prematurely or late (again, adjusting for potential biases associated with first phonemes). As for the English analyses, we used inverse RTs as the dependent variables, continuous centralised frequency and binary contrast-coded complexity (-0.5 = simple) as independent variables, and previous RT. Participants and items were included as random factors, and the effect of frequency was allowed to vary across participants. As the items were matched on the number of letters (as is common in studies on multi-letter rules; see Rastle & Coltheart, 1998; Rey et al., 1998), the simple ("garnir") condition had, by definition, more phonemes than the complex ("gâteau") condition. This also resulted in a lower number of syllables for complex (average = 2.0) compared to simple (average = 2.8) words. It was therefore decided, *a priori*, that the number of syllables should be included in the model, to act as a covariate.

Overall, there were no non-responses, and the accuracy rate was 97.5%. Accuracy was very high and evenly distributed across conditions (ranging from 95.7% for low-frequency simple words to 98.7% for high-frequency complex words).

An LME on the accuracy rates showed a main effect of frequency, $\beta = 0.5$, $z = 4.0$, $p < 0.0001$, reflecting higher accuracy for high- than low-frequency words. Neither the effect of complexity nor the complexity-by-frequency interaction reached significance, $p > 0.1$. This is likely to be reflect the overall high accuracy rates and lack of variability across conditions. For this reason, as for in Experiment 1, we draw conclusions from the RT data.

For the RT analyses, we removed one data point with RT < 300 ms, which yielded an approximately normal distribution of inverse RTs. When artificially dichotomising frequency (high: log frequency > 1; low: log frequency < 1), the average RTs are 608.5 ms (SD = 146.9) and 620.4 ms (SD = 156.4), respectively, for simple words, and 565.9 ms (SD = 114.5) and 603.5 ms (SD = 140.5), respectively, for complex words. Adjusting these means for the number of syllables yields, for simple words, 600.3 ms and 608.0 ms, for high- and low-frequency words respectively, and for complex words, 578.4 ms and 616.0 respectively.

The latency analyses showed a main effect of frequency, $\beta = -0.05$, $t = -5.3$, $p < 0.0001$. The main effect of the number of syllables, which was included as a covariate, was also significant, $\beta = 0.07$, $t = 9.8$, $p < 0.0001$. The main effect of complexity was not significant, $\beta = -0.01$, $t = -0.9$, $p = 0.4$, but the critical interaction between frequency and complexity was, indicating a steeper frequency slope for complex than simple words, $\beta = -0.06$, $t = 3.5$, $p = 0.0005$. The Bayes Factor provided strong evidence for the presence of this interaction, BF = 37.9 (±1.1%).

Follow-up analyses

An unexpected finding in Experiment 2 is the absence of a significant main effect of complexity. As the explanation of the complexity-by-frequency interaction, in the Orthographic Depth Hypothesis framework, is based on the assumption that

complex words are more difficult to process by the sublexical route than simple words, this finding might compromise our conclusion. A possible explanation is the inclusion of relatively high-frequency words in our item set. Previous research has shown that the complexity effect is diminished for high- compared to low-frequency words (Rey et al., 1998). LME provides the slope estimates at the point where the independent variables equal to zero. As we used centred log frequency as an independent variable, it is possible that the slope estimate of the complexity effect is based on a point where the frequency is too high to show a complexity effect. To test this possibility, we conducted follow-up tests of the effect of complexity separately for low-frequency (log frequency < 1) and high-frequency (log frequency > 1) words. Indeed, the data showed slower RTs for complex than simple items for low-frequency words, $\beta = 0.03$, $t = 1.6$, and faster RTs for complex than simple items for high-frequency words, $\beta = -0.05$, $t = -2.8$. The Bayes Factors provided equivocal evidence for the expected inhibitory complexity effect for low-frequency words, BF = 0.4 (±1.1%), and weak evidence for the unexpected facilitatory complexity effect for high-frequency words, BF = 4.9 (±0.9%).

As the facilitatory effect (faster RTs for complex than simple words) for high-frequency words goes both against the existing literature and existing models of reading, we considered possible confounds that could be driving this counter-intuitive pattern. In matching items with complex correspondences against items with simple correspondences, it is customary to match for the number of letters, not phonemes (Rastle & Coltheart, 1998; Rey et al., 1998). This is a conservative approach: As complex words contain more letters than phonemes, the complex condition necessarily has fewer phonemes than the simple condition. This could be counteracting the complexity effect in our analysis. Indeed, when adding the number

of phonemes as an additional predictor, we still get the predicted inhibitory effect (numerically) for low-frequency words, β = 0.03, t = 1.6 (adjusted means: 579.9 ms and 598.5 ms for complex and simple words, respectively), and the unexpected numerically facilitatory effect for high-frequency words, β = -0.04, t = -2.3 (adjusted means: 622.2 ms and 603.0 ms, for complex and simple words, respectively), but now the Bayes Factor provides weak evidence for the expected inhibitory effect for low-frequency words, BF = 3.9 (±0.8%), and equivocal evidence for the unexpected facilitatory effect for high-frequency words, BF = 1.6 (±0.8%). This means that the current data does not give us any conclusive evidence about whether or not there is a complexity effect for high-frequency words after taking into account the number of syllables and phonemes as a covariate, but suggests that there might be the expected inhibitory effect for low-frequency words. Note that in a post-hoc analysis of the full French data set which includes the number of phonemes as well as the number of syllables as covariates, we continue to get evidence for a frequency-by-complexity interaction, β = -0.07, t = -3.7, BF = 7.3 (±0.8%), suggesting that the key result is robust.

As with the English analyses, we performed one final post-hoc test to ensure that none of the potential covariates from Table 2 compromise our results. We repeated the analyses while including the main effect of OLD20 and PLD20 (which co-varied with complexity), and the main effect of bigram frequency and its interactions with frequency and complexity. For bigram frequency, missing values were replaced with the global mean. As in the previous model, we also included the number of phonemes, number of syllables, and the critical two main effects of frequency, complexity, and their interaction. Again, the pattern of results remained stable, with a main effect of frequency, β = -0.05, t = -5.1, an unexpected facilitatory

effect of complexity, β = -0.04, t = -2.4, and the critical interaction, β = -0.07, t = -3.5. The adjusted mean RTs are 574.2 ms and 604.3 ms for high-frequency complex and simple words, respectively, and 612.0 ms and 612.6 ms for low-frequency complex and simple words, respectively. The evidence for the critical interaction between complexity and frequency was BF = 13.6 (±0.86%).

In sum, we found evidence for the critical interaction, showing that the frequency effect is stronger for words with complex compared to words with simple correspondences. This suggests that, like unpredictability, complexity acts as a source of orthographic depth by impairing the sublexical route. This leads to a relative increase in the degree to which the lexical route contributes to the final output.

General Discussion

Although orthographic depth has been studied extensively throughout the past decades, it is unclear whether the complexity and the unpredictability of the sublexical correspondences affect skilled reading processes in the same way, or whether these two constructs have a differential effect on the cognitive processes (Schmalz et al., 2015). The Orthographic Depth Hypothesis proposes that in deep orthographies, the lexical route becomes relatively more important, because the sublexical information is less efficient in retrieving a correct pronunciation (Katz & Frost, 1992). We hypothesised that this may not be the case for orthographies with complex but predictable sublexical correspondences, such as French, because here the sublexical information is, in principle, sufficient to derive a correct pronunciation. However, increased lexical processing may be observed if complex correspondences slow down the sublexical route, thus allowing more time for the lexical route to retrieve the relevant phonological information. We found support for the latter possibility inasmuch as the frequency effect (a marker of lexical processing) was

greater for words with complex sublexical correspondences than for those with simple correspondences.

Predictability within models of reading

Experiment 1 indicates that, in a within-experiment manipulation, the frequency effect is stronger for English unpredictable ("ghost") than predictable ("forge") words. In line with the ODH (Katz & Frost, 1992) and with previous research (Frost et al., 1987), this suggest that unpredictability increases the relative reliance on lexical processing, as the sublexical processing cannot be resolved without lexical knowledge.

Note that, within a rule-based model of reading, the theoretical explanation of a predictability-by-frequency interaction is slightly different from what is likely to happen in the case of complexity, even though they result in an identical behavioural pattern (Coltheart et al., 2001). If the sublexical route uses a set of print-to-speech conversion rules, the sublexical output for words with irregular correspondences, which do not comply to the rules, will be an incorrect response (e.g., in English /dept/ instead of /dɛt/ for the written word *debt*). A conflict would then take place in the phonological buffer, when combining the output of the lexical and the sublexical routes. Such a conflict may be resolved by postponing the initiation of the verbal response, until sufficient activation from the lexical route has accumulated to trump the incorrect phonemic activation from the sublexical route. This would explain the main effect of unpredictability, because the pronunciation of unpredictable or irregular words is delayed, due to the conflict between the two routes. Furthermore, this conflict does not occur for words with predictable or regular correspondences, therefore the pronunciation does not need to be delayed until the lexical route trumps the activation of the sublexical route. As a result, relatively stronger lexical

involvement is needed to resolve the pronunciation of unpredictable words. For predictable words, the sublexical route does not need to be suppressed for a correct pronunciation.

Within a connectionist framework (Perry et al., 2007; Plaut et al., 1996), the sublexical route would be predicted to operate more slowly for unpredictable compared to predictable words. Unpredictable words, by definition, contain inconsistent correspondences (e.g., in the word "ghost", the grapheme *o* is inconsistent, as it can also be pronounced as in "lost"). It is possible that phonemic activation associated with inconsistent graphemes is slower than the activation of consistent graphemes (e.g., *sh* → /ʃ/ would be activated faster than *th* → /θ/). In this case, unpredictability (or, more specifically, inconsistency) would lead to an overall slow-down of the sublexical route, thus giving more time for the lexical or semantic information to contribute to the verbal output. Thus, in contrast to rule-based models, connectionist models would suggest that the mechanism responsible for the predictability-by-frequency interaction is very similar to the mechanism underlying the complexity-by-frequency interaction.

Complexity within models of reading

Experiment 2 examined whether the frequency effect would be stronger, for French, in words containing complex (multi-letter) correspondences ("gâteau"), compared to words with simple correspondences only ("garnir"). Again, there was evidence for an interaction, suggesting that complexity, like unpredictability, increases the relative importance of lexical processing.

As previous studies on the ODH have used cross-linguistic comparisons of pairs of orthographic systems that differed in both complexity and unpredictability (e.g., Serbo-Croatian/English, German/English), our study is the first to suggest that

complexity affects the ratio of lexical-to-sublexical processing. Presumably, this is due to a slow-down of the sublexical decoding process, which is caused by the application of complex multi-letter rules. More specifically, complex rules could lead to a conflict between the activation of the phoneme corresponding to a multi-letter grapheme and the phonemes corresponding to its underlying individual letters, as proposed by the dual-route cascaded (DRC) model of reading aloud (Rastle & Coltheart, 1998). In a word like "garnir", each letter maps onto its default phoneme. A simple word would lead to faster activation of the phonemes in the output buffer from the sublexical route, thus reducing the relative contribution of the lexical route in achieving the final pronunciation. For a word with complex correspondences, like "gâteau", the activation of the phonemes of the individual letters, *e* (→ /ɛ/), *a* (→ /a/), and *u* (→ /y/), would cause a conflict within the sublexical route, as the three letters need to be combined into a single grapheme and mapped onto the correct phoneme /o/. This would slow down the output of the sublexical route, such that the lexical route has a larger contribution to the final output.

We did not find a main effect of complexity, thus failing to replicate the results of Rastle and Coltheart (1998) and Rey et al. (1998). Including articulatory variables, namely the number of syllables and the number of phonemes, as covariates, provided a more coherent picture. Here, there was some evidence for an effect of complexity in the low-frequency condition, though this emerged only in the covariate analysis that included both the number of syllables and the number of phonemes. Thus, it seems that articulatory processes counteract the effect of complexity. Articulatory processes affect reading aloud latencies at a post-lexical stage (Cholin & Levelt, 2009; Cholin, Schiller, & Levelt, 2004), which results in in facilitation of the verbal response, driven by a smaller number of phonemes and syllables for all types

of words, regardless of frequency. The effect of complexity counteracts this facilitatory articulation-level effect especially for low-frequency words, as complexity operates on the sublexical level.

Notwithstanding the lack of a main effect of complexity, Experiment 2 provided strong evidence for an interaction between complexity and frequency in French. This suggests that sublexical information plays a role in determining the net ratio of lexical-to-sublexical processing, even if the output is driven to a great extent by the lexical route. While the process of reading aloud appears to happen at the same rate for complex as for simple words, there is relatively more contribution from the lexical than the sublexical route.

The frequency effect: Lexical, semantic, or sublexical marker?

Finally, it is worth expanding on our central assumption that the frequency effect is a marker of lexical processing. While it is generally assumed that frequency reflects some kind of threshold of the activation of entries in a mental orthographic lexicon (e.g., Coltheart et al., 2001; Taft, 1991), there are alternative views of how the frequency effect works. First, it is possible that frequency effects reflect other constructs that are strongly correlated, such as imageability (Strain, Patterson, & Seidenberg, 1995), age-of-acquisition (Zevin & Seidenberg, 2002), or contextual diversity (Adelman, Brown, & Quesada, 2006). We did not match for these variables, as it would have substantially limited the choice of items. The norms for these variables are not available for the majority of our items, therefore we can also not include them as *post-hoc* covariates in follow-up analyses. This does not present a problem for our conclusions, however, as these variables reflect lexical-semantic activation and thus measure processes which occur broadly within the lexical route. The DRC (Coltheart et al., 2001) and CDP+ (Perry et al., 2007; Perry, Ziegler, &

Zorzi, 2010) models make a distinction between an orthographic lexicon and a purely semantic route. The semantic route can be reached either by activation from the orthographic lexicon or the phonological lexicon, and in turn sends activation to the non-semantic lexical components. These models would therefore predict a close link between non-semantic and semantic lexical processes. Triangle models do not make a distinction between a semantic and a non-semantic route, as there is no purely orthographic representation of whole words, thus there is an even closer link between semantic marker effects and the lexical route (Plaut et al., 1996; Seidenberg & McClelland, 1989).

As a second alternative explanation of the frequency effect, it is possible that it reflects the frequency not of the whole word, but of the letters and letter clusters which are contained in the word. Thus, in a connectionist model, it is possible to show word frequency effects in the absence of an orthographic lexicon, because frequent letter clusters and their pronunciations are easier to learn (Plaut et al., 1996). This would imply that the frequency effect is a measure of sublexical processing. If a sublexical mechanism, reflecting the frequency of letter clusters, drives the interactions with complexity or predictability, one would expect the interaction to disappear once bigram frequency is taken into account. However, in Experiment 1 we found the frequency-by-predictability interaction while the manipulations did not co-vary with bigram frequency, and in Experiment 2, the frequency-by-consistency interaction remained robust after taking into account bigram frequency as a covariate. Thus, we can exclude the possibility that the frequency effect in our study reflects a sublexical process.

Conclusion

The current study is the first, to our knowledge, to empirically address the hypothesis that orthographic depth consists of various components that differentially affect skilled reading processes. The experiments reported here suggest that both complexity and unpredictability independently increase relative reliance on the lexical route. This provides support for the ODH, and the cognitive mechanism that Katz and Frost (1992) proposed as driving the cross-linguistic differences associated with orthographic depth: complexity and unpredictability both act to impair the efficiency of the sublexical route, which allows for a relatively greater influence of the lexical route in retrieving the word's pronunciation.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814-823.

Andrews, S. (1982). Phonological recoding: Is the regularity effect consistent? *Memory & Cognition, 10*(6), 565-575.

Andrews, S., & Scarratt, D. R. (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. doi:10.1037//0096-1523.24.4.1052

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics, 24*, 621-635.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412. doi:10.1016/j.jml.2007.12.005

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445-459. doi:10.3758/Bf03193014

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278. doi: 10.1016/J.Jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bridgeman, B. (1987). Is the Dual-Route Theory Possible in Phonetically Regular Languages. *Behavioral and Brain Sciences, 10*(2), 331-332.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*(5), 412.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and

improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Chateau, D., & Jared, D. (2003). Spelling–sound consistency effects in disyllabic word naming. *Journal of Memory and Language, 48*(2), 255-280.

Cholin, J., & Levelt, W. J. (2009). Effects of syllable preparation and syllable frequency in speech production: Further evidence for syllabic units at a post-lexical level. *Language and Cognitive Processes, 24*(5), 662-684.

Cholin, J., Schiller, N. O., & Levelt, W. J. (2004). The preparation of syllables in speech production. *Journal of Memory and Language, 50*(1), 47-61.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204-256. doi:10.1037//0033-295x.108.1.204

Cortese, M. J., & Simpson, G. B. (2000). Regularity effects in word naming: What are they? *Memory & Cognition, 28*(8), 1269-1276. doi:10.3758/Bf03211827

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., . . . Wagenmakers, E.-J. (2015). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, 1-8.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., . . . Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*(2), 488-496.

Forster, K. I., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instuments & Computers, 35*, 116-124.

Frith, U., Wimmer, H., & Landerl, K. (1998). Differences in Phonological Recoding in German- and English-Speaking Children. *Scientific Studies of Reading, 2*(1), 31-54.

Frost, R. (1994). Prelexical and Postlexical Strategies in Reading: Evidence from a Deep and Shallow Orthography. *Journal of Experimental Psychology: Learning, Memory & Cognition, 20*(1), 116-129.

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences, 35*(5), 263-279. doi:10.1017/S0140525x11001841

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for Visual Word Recognition and Orthographic Depth: A Multilingual Comparison. *Journal of Experimental Psychology: Human Perception & Performance, 13*(1), 104-115.

Glushko, R. (1979). The Organization and Activation of Orthographic Knowledge in Reading Aloud. *Journal of Experimental Psychology-Human Perception and Performance, 5*(4), 674-691.

Jared, D. (1997). Spelling-Sound Consistency Affects the Naming of High-Frequency Words. *Journal of Memory and Language, 36*(4), 505-529. doi:10.1006/jmla.1997.2496

Jared, D. (2002). Spelling-Sound Consistency and Regularity Effects in Word Naming. *Journal of Memory and Language, 46*(4), 723-750. doi:10.1006/jmla.2001.2827

Jared, D., McRae, K., & Seidenberg, M. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language, 29*(6), 687-715. doi:10.1016/0749-596X(90)90044-Z

Katz, L., & Feldman, L. (1983). Relation between pronunciation and recognition of printed words in deep and shallow orthographies. *Journal of Experimental Psychology-Learning Memory and Cognition, 9*(1), 157-166. doi:10.1037/0278-7393.9.1.157

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* (pp. 67-84). Amsterdam: Elsevier Science Publishers.

Kearns, D. M., Steacy, L. M., Compton, D. L., Gilbert, J. K., Goodwin, A. P., Cho, E., . . . Collins, A. A. (2014). Modeling Polymorphemic Word Recognition Exploring Differences Among Children With Early-Emerging and Late-Emerging Word Reading Difficulty. *Journal of Learning Disabilities*, 0022219414554229.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Language Sciences, 1*, 174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English

words. *Behavior Research Methods, 44*(1), 287-304. doi:10.3758/S13428-011-0118-4

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition, 63*, 315-334.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers, 36*(1), 156-166.

Mann, V., & Wimmer, H. (2002). Phoneme awareness and pathways into literacy: A comparison of German and American children. *Reading and Writing: An Interdisciplinary Journal, 15*, 653-682.

Marinus, E., & de Jong, P. F. (2010). Variability in the word-reading performance of dyslexic readers: Effects of letter length, phoneme length and digraph presence. *Cortex, 46*(10), 1259-1271. doi:10.1016/j.cortex.2010.06.005

Marinus, E., Nation, K., & de Jong, P. (2015). Density and length in the neighbourhood: Explaining cross-linguistic differences in learning to read in English and Dutch. *Journal of Experimental Child Psychology, 139*, 127-147. doi:10.1016/j.jecp.2015.05.006

Medler, D. A., & Binder, R. J. (2005). MCWord: An On-Line Orthographic Database of the English Language. Retrieved from http://www.neuro.mcw.edu/mcword/

Morey, R. D., & Rouder, J. N. (2014). Package "BayesFactor". Retrieved from http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28*(04), 661-677.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers, 36*(3), 516-524.

Perry, C., Ziegler, J., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychol Rev, 114*(2), 273-315. doi:10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology, 61*(2), 106-151.

Plaut, D. C., McClelland, J. L., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115. doi:10.1037/0033-295x.103.1.56

Pritchard, S. C., Coltheart, M., Palethorpe, S., & Castles, A. (2012). Nonword Reading: Comparing Dual-Route Cascaded and Connectionist Dual-Process Models With Human Data. *Journal of Experimental Psychology-Human Perception and Performance, 38*(5), 1268-1288. doi:10.1037/A0026703

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods, 39*(4), 859-862. doi:10.3758/bf03192979

Rastle, K., & Coltheart, M. (1998). Whammies and double whammies: The effect of length on nonword reading. *Psychonomic Bulletin & Review, 5*(2), 277-282.

Rau, A. K., Moll, K., Snowling, M. J., & Landerl, K. (2015). Effects of orthographic consistency on eye movement behavior: German and English children and adults process the same words differently. *Journal of Experimental Child Psychology, 130*, 92-105.

Rey, A., Jacobs, A. M., Schmidt-Weigand, F., & Ziegler, J. C. (1998). A phoneme effect in visual word recognition. *Cognition, 68*(3), B71-B80. doi:10.1016/S0010-0277(98)00051-1

Robidoux, S., & Pritchard, S. C. (2014). Hierarchical clustering analysis of reading aloud data: a new technique for evaluating the performance of computational models. *Frontiers in Psychology, 5*. doi:10.3389/Fpsyg.2014.00267

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. doi:10.3758/Pbr.16.2.225

Schmalz, X., Marinus, E., Coltheart, M., & Castles, A. (2015). Getting to the bottom of orthographic depth. *Psychonomic Bulletin and Review, 22*(6), 1614-1629. doi:10.3758/s13423-015-0835-2

Schmalz, X., Marinus, E., Robidoux, S., Palethorpe, S., Castles, A., & Coltheart, M. (2014). Quantifying the reliance on different sublexical correspondences in German and English. *Journal of Cognitive Psychology, 26*(8), 831-852.

Seidenberg, M., & McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review, 96*(4), 523-568.

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

Share, D. (2008). On the Anglocentricities of Current Reading Research and Practice: The Perils of Overreliance on an "Outlier" Orthography. *Psychological Bulletin, 134*(4), 584-615.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 0956797611417632.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1140.

Taft, M. (1991). *Reading and the Mental Lexicon*. Hove: Lawrence Erlbaum.

Treiman, R., Kessler, B., & Bick, S. (2003). Influence of consonantal context on the pronunciation of vowels: A comparison of human readers and computational models. *Cognition, 88*(1), 49-78. doi:10.1016/s0010-0277(03)00003-9

Turvey, M., Feldman, L., & Lukatela, G. (1984). The Serbo-Croatian orthography constrains the reader to a phonologically analytic strategy. In L. Henderson (Ed.), *Orthographies and reading: Perspectives from cognitive psychology, neuropsychology, and linguistics* (pp. 81-89). Hillsdale, NJ: Lawrence Erlbaum.

van den Bosch, A., Content, A., Daelemans, W., & de Gelder, B. (1994). Measuring the complexity of writing systems. *Journal of Quantitative Linguistics, 1*(3), 178-188.

van Heuven, W., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60*(4), 502-529.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971-979. doi:Doi 10.3738/Pbr.15.5.971

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47*(1), 1-29.

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

Ziegler, J., Perry, C., & Coltheart, M. (2003). Speed of lexical and nonlexical processing in French: The case of the regularity effect. *Psychological Bulletin & Review, 10*(4), 947-953.

Acknowledgements

We thank Melvin Yap and David Balota for making the trial-level ELP data available to us, and Sachiko Kinoshita for valuable discussions.

Correspondence concerning this manuscript should be addressed to X. Schmalz, email: xenia.schmalz@gmail.com, phone: +39 320 639 2523, fax: +39 049 827 6547; Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università degli Studi di Padova, Via Venezia 15, 35131 Padova, Italy.

Table 1:

Potential moderators, and how they co-vary with the critical manipulations of Experiment 1.

Potential covariate	**Main effect of frequency**	**Main effect of predictability**	**Interaction of predictability and frequency**	**Overall average and standard deviation**
Number of letters	t = -1.51, p = 0.13	t = -0.94, p = 0.35	t = -0.91, p = 0.36	4.99 (0.93)
Number of phonemes	t = -1.58, p = 0.12	t = 1.52, p = 0.13	t = -0.26, p = 0.79	3.78 (0.82)
Orthographic N	t = 2.02, p = 0.04 *	t = 1.04, p = 0.30	t = -0.23, p = 0.82	5.82 (4.35)
Bigram frequency	t = 0.49, p = 0.62	t = 0.12, p = 0.91	t = -0.90, p = 0.37	68.91 (73.76)
Ratio of letters to phonemes	t = 0.55, p = 0.58	t = -3.10, p < 0.01*	t = -0.80, p = 0.42	1.36 (0.32)
Phonological Levenshtein Distance	t = -3.80, p < 0.01 *	t = -0.06, p = 0.95	t = -0.74, p = 0.46	1.42 (0.31)

Note: Frequency counts are based on the English subtitle corpus (New et al., 2007); orthographic N counts are retrieved from the British Lexicon Project (Keuleers et al., 2012); bigram frequency is from the MCWord database (Medler & Binder, 2005); Phonological Levenshtein Distance is retrieved from the English Lexicon Project (Balota et al., 2007).

Table 2:

Potential moderators, and how they co-vary with the critical manipulations of Experiment 2.

Potential covariate	**Main effect of frequency**	**Main effect of complexity**	**Interaction of complexity and frequency**	**Overall average and standard deviation**
Number of letters	t = -0.12, p = 0.91	t = 1.59, p = 0.11	t = 1.59, p = 0.11	6.62 (1.60)
Number of phonemes	t = -0.93, p = 0.35	t = -13.22, p < 0.01 *	t = 0.99, p = 0.32	5.54 (1.70)
Orthographic Levenshtein Distance	t = -1.58, p = 0.12	t = -2.76, p < 0.01 *	t = -0.71, p = 0.48	2.03 (0.47)
Bigram frequency	t = 1.13, p = 0.26	t = 5.57, p < 0.01 *	t = -3.42, p < 0.01 *	1128.61 (603.59)
Grapheme consistency	t = -1.24, p = 0.22	t = -0.32, p = 0.75	t = 0.63, p = 0.53	84.14 (9.44)
Phonological Levenshtein Distance	t = -0.59, p = 0.56	t = -7.79, p < 0.01 *	t = -0.76, p = 0.45	1.82 (0.60)

Note: Log subtitle frequency, Orthographic Levenshtein Distance, and Phonologicial Levenshtein Distance are retrieved from Lexique (New et al., 2004); grapheme consistency and bigram frequency from Manulex (Lété et al., 2004). Note that Manulex has the bigram frequency for only 314 out of the 384 words; missing cells were excluded for the analysis which included bigram frequency as the dependent variable.