Methodological and theoretical issues in cross-linguistic reading research

Xenia Schmalz, BSc (Psyc) (Hon)

Department of Cognitive Science

& ARC Centre of Excellence in Cognition and its Disorders

Macquarie University

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Cognitive Science at Macquarie University.

Table of contents

Thesis summary

In this thesis, we explore methodological and theoretical issues associated with the concept of orthographic depth. In the first section (Papers 1 - 5), we conducted a series of word and nonword reading experiments. We compare the sublexical correspondences underlying reading in English, which is considered to be a deep orthography, and German, which is considered to be a shallow orthography. In experiments with adults, we aimed find a sensitive and reliable experimental manipulation to explore the reliance on different types of sublexical correspondences across languages. We follow up with experiments with children, to assess the developmental trajectory of sublexical processing. In the second section (Paper 6), we discuss issues with defining the concept of orthographic depth, and provide some suggestions as to how this concept can be quantified on a linguistic level.

Statements and signature

I certify that the work in this thesis entitled “Methodological and theoretical issues in cross-linguistic reading research” has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself has been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in the thesis. The research presented in this thesis was approved by the Macquarie University Faculty of Human Sciences (FHS) Ethics Committee (reference number: 5201200053), and the Potsdam University Ethics Committee (reference number: 4/2013).

Signed

Xenia Schmalz

Acknowledgements

First and foremost, I want to thank my supervisors, Eva Marinus, Max Coltheart, and Anne Castles. I could not have wished for more knowledgeable and supportive supervisors, who have gone beyond their responsibilities in sharing their knowledge of the field of reading research, as well as approaches to science in general.

I am also grateful to Reinhold Kliegl and his lab, who hosted me during my long-term stay at Potsdam University, and especially to Petra Schienmann, whose help with recruiting and testing participants in Potsdam was indispensable.

The optimisation procedure from Papers 4 and 5 was implemented by Serje Robidoux, who has also guided me through the data analyses for Paper 2. I would like to thank him for his important contributions to the thesis. I am grateful to Sachiko Kinoshita, who was always happy to provide helpful advise, and introduced me to Bayes Factor testing, which allowed me to draw conclusions from some otherwise uninterpretable results. Linda Buckley and Sallyanne Palethorpe transcribed the English nonword responses which are reported in this thesis.

The thesis has benefited immensely from feedback during various conferences and lab visits. I would like to thank Reinhold Kliegl, Heinz Wimmer, and Becky Treiman, for hosting me in their respective labs, and Anne Castles, who provided funding for me to present some of my work at ESCOP (Budapest, 2013).

Thanks are extended to family and friends for their support throughout my candidature, especially to Wei, Leidy, and Bobi, for words of encouragement, and distractions, when needed.

The thesis is dedicated to Oma Maria Schmalz, who would have appreciated the hard work that went into its preparation.

General Introduction

General Introduction

Reading is a complex cognitive phenomenon: it requires the coordination of numerous subskills that need to work in concert to integrate information about a word's visual features, orthography, phonology, and semantics. This complexity makes the study of reading an interesting challenge in cognitive psychology. In addition to its theoretical interest, reading also forms the basis of everyday life, and is an important skill in Western society. This provides a strong practical motive for studying its underlying cognitive processes, as understanding the details of how reading works will, in the long term, help with creating effective classroom instructions and remediation programs for individuals who struggle with acquiring this highly complex skill.

Although theories and models of reading are arguably the most-developed and well-specified of any area in cognitive psychology, many questions remain about its underlying processes. One way of filling in some of the gaps is to study reading across languages: establishing to what degree processes underlying reading differ helps us to unveil how universal cognitive competencies interact with language characteristics. This is relevant to a broad range of issues, because reading development, dyslexia, and skilled reading all occur within the context of a given orthography.

In this thesis, we present a series of papers addressing open questions in cross-linguistic reading research. The experiments in this thesis focus mainly on similarities and differences in cognitive processes underlying single-word reading in German and English. These two orthographies have been used in previous research in cross-linguistic comparisons because they differ from each other in terms of their orthographic depth, while being comparable in terms of other characteristics, such as the presence of complex onset clusters (e.g., Seymour, Aro, & Erskine, 2003) and relatively high prevalence of monosyllabic words (e.g., Ziegler, Perry, & Coltheart, 2000). Orthographic depth,

broadly speaking, refers to the reliability of the relationship between letters and sounds. In Papers 1-5, we aim to provide further insights into how the relationship between print and speech in the orthography affects the mechanisms that are used during reading. An intuitive prediction (which has been backed up by previous work) is that the print-to-speech reliability affects the cognitive mechanisms that are used for print-speech conversion (Frith, Wimmer, & Landerl, 1998; Landerl, Wimmer, & Frith, 1997; Ziegler, Perry, Jacobs, & Braun, 2001; Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003). Here, we aim to further specify the mechanisms that are different during reading in the two orthographies (Papers 1 - 5). In addition, we explore whether the print-to-speech reliability also affects the development of lexical processes (Paper 5).

In Paper 6, we discuss theoretical issues with defining orthographic depth, and what particular language-level aspects are likely to drive behavioural differences as a function of orthographic depth that have been found throughout the thesis, and by previous research. Finally, in the general conclusion, we attempt to bring together the theoretical and empirical issues encountered throughout the thesis, and discuss the implications for theories of skilled reading and reading acquisition, cross-linguistic theories of reading, and future research.

Theories of skilled reading in English and German

The thesis was conducted broadly within a dual-route framework. According to dual route theories, reading occurs via two cognitive mechanisms that operate in parallel: the first mechanism is a lexical look-up procedure, and the second uses knowledge of the statistical regularities between letters and sounds to compute the pronunciation. Two major computational models of the dual-route model are the Dual Route Cascaded (DRC; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) and the Connectionist Dual Process (CDP) +/++ (Perry, Ziegler, & Zorzi, 2007; Perry, Ziegler, & Zorzi, 2010) models. The

original models were implemented in English, but German versions exist of both the DRC (Ziegler et al., 2000) and the CDP+ (Perry, Ziegler, Braun, & Zorzi, 2010).

The dual-route framework was originally developed by English-speaking researchers on the basis of work with acquired dyslexics showing a functional dissociation between the two procedures (Marshall & Newcombe, 1973). Some patients, following brain damage, showed a selective deficit in reading unfamiliar words or nonwords, which suggests that their sublexical route was damaged, as their familiar word reading skills were intact. Other patients had intact nonword reading skills, but showed deficits in reading aloud so-called *irregular words*, or words which are not predictable based on the orthography's print-to-speech regularities (e.g., *yacht*, *colonel*). This pattern suggests that the sublexical route is intact: it allows the patients to read aloud not only nonwords, but also any word that can be deciphered correctly by the sublexical route. For irregular words like *yacht*, however, the lexical route is required, because the sublexical route will attempt to decode it by print-speech correspondences and provide a "regularised" output ("/jætʃt/"). This double dissociation motivated the dual-route framework.

Dual-route models are often contrasted single-route models (Glushko, 1979; Kay & Marcel, 1981; Seidenberg & McClelland, 1989). Single-route models propose that a single mechanism is sufficient to derive the pronunciation of both nonwords and irregular words. In a computational implementation, the triangle (PDP) models propose a learning algorithm, which can derive the regularities between print and speech (Harm & Seidenberg, 1999; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). As a result, the model can generalise this print-to-speech correspondence knowledge to pronounce unfamiliar words or nonwords. The model also develops sensitivity to regularities that exist when larger units are taken into account,

such that given a sufficient amount of training, connections between orthographic whole words and their phonological forms can be established. This allows a single-route mechanism to read aloud irregular words as well as nonwords.

Nevertheless, PDP models have a dual-route structure, in the sense that the pronunciation of a word can be computed either via this single-route mechanism (the orthography → phonology pathway), or by lexical access via semantics (the orthography → semantics → phonology pathway). As dual-route models, such as the DRC or CDP+/++, also have a lexical-semantic pathway, the difference between dual-route and triangle models becomes that the dual-route makes a distinction between sublexical and lexical print-to-speech conversion, while the triangle models do not. This difference between the models is not relevant to the current thesis: when we refer to lexical processing, we cannot distinguish between a process that happens in a non-semantic lexical route or one which is mediated by semantics. The aim of the present thesis was not to adjudicate between these two classes of model but, where relevant, different interpretations of particular findings in the context of these models are provided.

Initially, some researchers pointed out that the dual-route framework was "anglocentric", as the concept of irregular words does not apply to the same extent to other orthographies (Bridgeman, 1987; Turvey, Feldman, & Lukatela, 1984). Specifically, the concept of irregular words was argued to be specific to English, because the high number of irregular words in English is a result of the somewhat special characteristics of the orthography. Other alphabetic orthographies have a closer relationship between print and speech, meaning that there are fewer cases where the pronunciation of a given word is unpredictable. This concept is referred to as *Orthographic Depth*: an orthography with a close relationship between print and speech is called "shallow", and an orthography where these sublexical correspondences are more

ambiguous is called "deep". Compared to all other European orthographies, English is considered to be an outlier in terms of its depth. Therefore, the presence of irregular words is specifically prominent in English.

Due to the special characteristics of English, a strong view was put forward, namely that in shallow orthographies, which have a one-to-one correspondence between letters and sounds, there is no pressure for the lexical route to develop at all (Bridgeman, 1987; Turvey et al., 1984). A widely-cited example of a shallow orthography is the case of the Serbo-Croatian orthography: here, it was argued, every word is pronounced the way it is spelled and can therefore be decoded sublexically.

Subsequent research has shown that such a view is not tenable (for reviews, see Besner & Smith, 1992; Katz & Frost, 1992). Upon close investigation of this issue, it became clear that each orthography requires both some degree of lexical processing, and some degree of sublexical processing: there are no orthographies which are perfectly shallow (i.e., where all words have a one-to-one correspondence between letters and sounds), and no orthographies which are completely deep (i.e., the relationship between print and speech is completely arbitrary). Even Serbo-Croatian, despite being relatively shallow, requires lexical knowledge for lexical stress assignment. As well, there is evidence that lexical processing occurs in shallow orthographies and sublexical processing in deep ones. Studies in Chinese, where written words often have an abstract relationship to their pronunciation, have found early activation of phonology (Tan & Perfetti, 1998). Studies in shallow orthographies, such as Italian, have shown the involvement of lexical processing. A report of semantic priming effects indicates early and automatic activation of the lexical-semantic system during single-word reading (Tabossi & Laghi, 1992). Furthermore, cases of Italian surface dyslexics have been reported, where individuals show deficient reliance on lexical information compared to

controls: this is manifested as problems with distinguishing homophones, such as *lago* - lake and *l'ago* - the needle. (Job, Satori, Masterson, & Coltheart, 1984; Zoccolotti, De Luca, Di Pace, Judica, & Orlandi, 1999). This results from over-reliance on a sublexical conversion mechanism rather than whole-word knowledge compared to controls.

Therefore, it is now generally acknowledged that reading in any orthography relies on both a lexical and a sublexical procedure (e.g., Katz & Frost, 1992; Kuo et al., 2004; Share, 2008; Tabossi & Laghi, 1992; Ziegler et al., 2000). Still, a great amount of data indicates that there are differences in reading across orthographies: even just for the comparison of German and English, behavioural studies have shown differences in various aspects of reading, ranging from speed of development (Frith et al., 1998; Landerl et al., 1997; Seymour et al., 2003; Wimmer & Goswami, 1994), through to characteristics of reading disabilities (Landerl et al., 2013; Wimmer, Mayringer, & Landerl, 2000), and to the cognitive processes underlying skilled reading (Rau, Moll, Snowling, & Landerl, 2015; Schmalz et al., 2014, i.e., Paper 4; Ziegler et al., 2001).

The most common explanation for the behavioural differences between English and German is that the two orthographies differ in terms of orthographic depth, as English is considered to be a deep orthography, and German is considered to be shallow (Borgwaldt, Hellwig, & de Groot, 2005; Seymour et al., 2003). Below we provide an overview of two theories of orthographic depth that make predictions about how cognitive processing should differ in a pair of orthographies that (arguably) represent two points on opposite ends of the depth continuum, such as German and English.

The Orthographic Depth Hypothesis

According to the Orthographic Depth Hypothesis (ODH; as advanced by Frost, Katz, & Bentin, 1987; Katz & Frost, 1992) orthographic depth varies as a continuum across the world's orthographies, and although readers of every orthography rely both on

a lexical and a sublexical procedure, the relative degree of lexical-to-sublexical processing depends on orthographic depth. In this view, the cognitive mechanisms that underlie reading as a function of orthographic depth are the same, but the degree to which they are recruited differs. We can therefore describe the ODH as proposing that orthographic depth causes quantitative cross-linguistic differences in cognitive processing underlying reading.

The ODH proposes the following driving mechanism for the cross-linguistic differences: Orthographic depth reflects the ease or difficulty with which the sublexical route can derive the correct pronunciation. In deep orthographies, the sublexical route is unreliable (as in the case of English where, for example, the letter *g* is pronounced differently in the words *gist* and *gift*) or incomplete (as in the case of unpointed Hebrew, where vowel information is not represented), or the sublexical correspondences are complex (such as the presence of multi-letter rules, e.g., the five-letter grapheme *aient* being pronounced as "/ɛ/" in the French word *étaient*). The *inconsistency*, *incompleteness*, or *complexity* of the sublexical correspondences is proposed to slow down the functioning of the sublexical route. As the lexical and sublexical route work in parallel, the slowing-down of the sublexical route should give more chances for the lexical route to derive the correct pronunciation. Direct evidence for this comes from a study with Hebrew, where the completeness of the script can be manipulated by either excluding vowel markings or including them. In such an experiment, Frost (1994) found stronger lexical marker effects when they used the same words without vowel markings, compared to the full and complete script.

The Psycholinguistic Grain Size Theory

A more recent theory of reading across orthographies is the Psycholinguistic Grain Size Theory (PGST; Ziegler & Goswami, 2005). The authors of the PGST propose

that "orthographic consistency may affect not so much the relative contribution of phonology (i.e., the specific mix orthographic [lexical] and phonological [sublexical] pathways), but rather the very nature of the phonological [sublexical] processes themselves" (p. 379, Ziegler et al., 2001). This is based on the observation that simple correspondences, which are mostly sufficient for print-to-speech conversion in shallow orthographies, are unreliable in deep orthographies (Peereman & Content, 1998; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). Therefore, different types of correspondences (e.g., bodies, syllables) are required to drive the sublexical conversion process in deep orthographies. In contrast to the ODH, we can then describe the PGST as proposing a qualitative change in the nature of sublexical processing, rather than a quantitative change in the ratio of lexical-to-sublexical processing.

The empirical support for the PGST focuses on the use of body-rime correspondences, where bodies are the vowel and (optional) final consonant of a monosyllabic word (e.g., *-orld* for the word *world*), and the rime is its phonological equivalent ("/ɜːld/"). English readers are proposed to rely to a greater extent on bodies, because in English, bodies are more predictive of the pronunciation than single letters or graphemes (Peereman & Content, 1998; Treiman et al., 1995). Ziegler and colleagues measured reliance on bodies with a body-N manipulation in both children and adults (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). Words and nonwords with many body neighbours (i.e., words with the same body, such as *at*, *hat*, and *brat*, which are all body neighbours of the word *cat*) were found to have faster reading aloud latencies (and higher accuracy, for children) than items with few body neighbours (e.g., the word *jazz*, which has no body neighbours). According to the studies of Ziegler (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003)(2001 & 2003), this effect is stronger in English than in German (but see also Papers 2 and 3 of the current thesis).

In the context of the dual-route framework, the PGST indicates that the types of sublexical correspondences that are stored in an English reader's sublexical route are different from those stored in a German reader's sublexical route. It should be noted that in the later publications on the PGST (Goswami & Ziegler, 2006; Ziegler & Goswami, 2005, 2006), the authors emphasised that their theory is not compatible with the dual-route framework: rather than drawing a clear distinction between lexical and sublexical processing, the authors propose a continuum. According to this view, letter-sound correspondences form the smallest units of the reading system, larger orthographic units and their phonological equivalents as intermediate grain-sizes, and whole words, and possibly word combinations, as largest grain-sizes (Goswami & Ziegler, 2006; Ziegler & Goswami, 2005, 2006). However, a common prediction of both views relates to greater reliance on body-rime correspondences and other sublexical clusters in English than in German, and therefore we focus on the use of these throughout the thesis (see Papers 1-3 about reliance on body-rime correspondences, and Papers 4-5 on the use of context-sensitive correspondences).

Quantitative and qualitative differences underlying reading in English and German

In summary, there are two theories of reading that propose cross-linguistic differences in reading between English and German, due to their differences in orthographic depth. Within the framework of the dual-route model, these can be interpreted as follows: the ODH predicts that the ease with which the sublexical route can assemble the correct pronunciation influences the ratio of lexical-to-sublexical processing. In English, compared to German, the sublexical route is characterised both by its complexity and its lack of predictability. English contains many complex sublexical rules compared to German (e.g., in the English implementation of the dual-route cascaded model, *sch* is pronounced as "/ʃ/" when it occurs in the first position of a word and is

followed by a vowel; in the German version of the dual-route cascaded model, *sch* is pronounced as "/ʃ/" in all positions and contexts; Coltheart et al., 2001; Ziegler et al., 2000). As well, relative to German, the application of English rules is unpredictable in that it results in a failure to read correctly a relatively large percentage of words (Ziegler et al., 2000). Both the complexity and the predictability of the English sublexical system compared to German should result in a stronger relative influence of the lexical route. We address this hypothesis in Papers 1 (adults) and Paper 5 (children). In particular, we attempt to establish to what degree the reading of irregular words differs in English and German.

Taken within the framework of the dual-route model, the PGST proposes that the nature of sublexical processing differs as a function of orthographic depth (Ziegler et al., 2001). In Paper 1, we follow up on the original reports of stronger reliance on bodies in English than German using a different task than that used by Ziegler and colleagues (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). In Paper 2, we report on a large-scale analysis of Body N effects in 8 experiments that used the same paradigm as Ziegler et al. (2001 & 2003); in Paper 3, we aim to replicate their results with children. In Papers 4 and 5, we present a nonword reading aloud task, and an optimisation approach that allows us to quantify the degree to which German and English readers rely on different types of different sublexical correspondences, based on their nonword pronunciations.

Theories of reading acquisition in English and German

Acquisition of the sublexical route

Within the developmental literature on cross-linguistic differences, there is one major finding: namely, that it is easier to learn to read in shallow compared to deep orthographies. This has been demonstrated consistently in relation to German and English

(Aro & Wimmer, 2003; Frith et al., 1998; Landerl, 2000; Seymour et al., 2003; Wimmer & Goswami, 1994).

This finding fits within the framework of the PGST: as the sublexical regularities underlying reading in a deep orthography are more complicated, they take a longer time to acquire. This also makes some predictions about the developmental trajectory of the influence of various sublexical units for both English and German: if simpler orthographic units and correspondences are easier to learn, then these should be the dominant drivers underlying the sublexical decoding process for younger readers. With increasing reading experience, children should also acquire the subtler and more complex regularities between print and speech, so more complex statistical regularities, such as body-rime correspondences, should be used as pronunciation heuristics only after a considerable amount of reading experience.

Although this prediction is not made explicitly by the PGST, it is empirically testable. It is also interesting, because it is relevant to another debate in the developmental literature: namely the large versus small units first debate. According to large-units-first theories, children start learning to read via "large" units, such as bodies, as the phonological awareness of these develops before the phonological awareness of "small" units or phonemes (Goswami, 1993, 2002; Goswami & Bryant, 1990). Small-units first theories argue that children start learning to read by reliance on "small" units, because awareness of phonemes (rather than larger phonological units) consistently emerges as a strong predictor of reading acquisition (Duncan, Seymour, & Hill, 1997; Hulme et al., 2002).

We address this question, whether reliance on large units (i.e., bodies or context-sensitive GPC rules) increases or decreases across reading acquisition, in Papers 3 and 5, where we use a body-N and a nonword reading aloud paradigm respectively. The large-

versus-small-units debate originates from English-speaking countries; therefore, we can make no predictions about how the English findings will compare to findings with German children. If we do find any cross-linguistic differences in the reliance on large units for children, it is not clear whether these will emerge in the beginning stages of reading acquisition, or whether it takes time for the different nature of the print-speech regularities to show an effect on the sublexical mechanisms that are used by children.

Acquisition of the lexical route

Earlier versions of the PGST make no predictions about the development of the lexical route, as the proposed cross-linguistic differences centre on the functioning of the sublexical route (Ziegler et al., 2001). If we consider the later proposals of the PGST (Goswami & Ziegler, 2006; Ziegler & Goswami, 2005, 2006), we would treat words as "large units". On the surface, this might lead us to predict that these large units should be more important in English than German, because large units should be more important in deep than in shallow orthographies. Such a claim could also be said to be consistent with the ODH[1]: the sublexical regularities underlying English are more complex and less reliable in English compared to German, which might push for stronger reliance on lexical processes in English.

Upon closer inspection, however, it is not clear that such a view is justified, if we consider the strong inter-dependent nature of lexical and sublexical processes in learning to read. Initial knowledge of letters and their corresponding sounds allows the children to decode new words, which serves as a powerful self-teaching mechanism: once words have been encountered and decoded, they can be stored in the mental lexicon as orthographic whole-word forms. Once the child starts to build up a comprehensive store

[1] The reason that it is not is that the ODH is a theory of skilled reading, and makes no predictions about reading acquisition.

of orthographic entries, the sublexical knowledge relating to the more subtle regularities can be further refined (Share, 1995; Ziegler, Perry, & Zorzi, 2014). Therefore, the development of lexical and sublexical processes is fundamentally intertwined. Given this view, we would expect that orthographic depth would affect both lexical and sublexical processes: without a sound functioning of the sublexical route, children should find it more difficult to establish lexical entries.

We examine this prediction in Paper 5, where German and English children read irregular words. Although previous studies have established that for English compared to German children, performance is poorer on both word and nonword reading, to our knowledge, no previous study has used an irregular word reading task. If the words in the experimental set are regular, they can, in theory, be decoded by the sublexical route. Therefore, an experiment that does not use irregular words does not allow us to make any strong conclusions about the efficiency of the lexical route.

Summary

In the current thesis, we aimed to understand why and how reading in German and English might be different. Previous work has shown behavioural differences between reading in English and reading in German, and attributed these to orthographic depth. As orthographic depth relates to the regularities between print and speech in the orthography, most studies are concerned with establishing the exact mechanisms that are used by the cognitive system to apply print-to-speech correspondences during reading (i.e., the sublexical route). We conducted experiments to understand the sublexical and lexical mechanisms that underlie cognitive processing in the two orthographies in both adults and children. In the General Discussion, we assess the implications of the results for models of skilled reading and reading acquisition, methodological issues that were encountered throughout, and possible solutions to these issues.

Paper 1: Consistency and Regularity Effects in German and English

Abstract

This paper aims to assess differences and similarities in the size of the consistency and regularity effects in English and German. Consistency relates to the print-speech correspondence reliability of bodies (large units), while regularity relates to the reliability of graphemes (small units). As the English orthography contains more words with unreliable print-speech correspondences than German, we seek to establish whether English readers process these in the same way as German readers. We find no cross-linguistic differences in the size of the consistency effect, suggesting that the reliance on bodies is similar in English and German. In terms of the regularity effect, the differences in the nature of the grapheme-phoneme correspondence rules that are used to define regularity prevent a direct comparison. We conduct follow-up experiments and discuss issues with defining these rules in the first place.

1.1. Consistency and Regularity Effects in German and English

Skilled reading requires the automatic activation of orthography, which can then be mapped onto a corresponding phonological representation. During this process, sublexical units, meaning letter or letter clusters without lexical information, are used in addition to whole-word knowledge. In everyday reading, this sublexical process is particularly useful for reading unfamiliar words: these cannot benefit from direct lexical access, as they have no entry in the mental lexicon. Understanding the details of how this route operates has theoretical implications, as models of single-word reading aloud make fundamentally different assumptions about how sublexical processing works (M. Coltheart et al., 2001; Perry et al., 2007; Plaut et al., 1996). The current study is specifically concerned with the use of orthographic units called graphemes ("small" units) and bodies ("large" units), how these can be defined, and whether their use can be compared directly across orthographies.

The nature of sublexical decoding has been proposed to differ across languages. According to a leading cross-linguistic theory of reading, the Psycholinguistic Grain Size Theory (PGST; Ziegler & Goswami, 2005), a concept called orthographic depth is a driving factor in determining the nature of the sublexical decoding process. An orthography is considered to be shallow when the correspondence between letters and sounds approaches a one-to-one relationship, and deep when this relationship is ambiguous or opaque. For example, English is considered to be a deep orthography, due to the presence of many words with ambiguous print-speech mappings, such as the words *colonel* or *yacht*.

The PGST proposes that in deep orthographies, readers are forced to develop sensitivity to "large"[2] units in an attempt to reduce the uncertainty associated with the pronunciation of an unfamiliar word. These include bodies and syllables. Here, we focus in particular on the concept of bodies, since these have been the focus of previous empirical investigation (Ziegler & Goswami, 2005). For a monosyllabic word, a body is the orthographic unit that consists of the vowel and the (optional) succeeding consonants, such as *-oard* for the word *board*. The rime is the phonological equivalent of the body ("/oːd/" for the body *-oard*).

Bodies are hypothesised to be especially important for reading in deep orthographies, because they tend to be more predictive of a word's pronunciation than letters or graphemes[3], such as the body *-alm* in the word *psalm*: if a reader is unfamiliar with this word, the knowledge that the body *-alm* maps onto the rime "/ɐːm/" is more helpful for deriving the correct pronunciation than knowledge of the mappings of the individual graphemes *a*, *l*, and *m* ("/ælm/"). The statistical advantage of bodies over graphemes has been shown through corpus analyses of English monosyllabic words (Peereman & Content, 1998; Treiman et al., 1995). Conversely, readers of orthographies with unambiguous letter-to-sound correspondences (so-called shallow orthographies) can rely on "small" units such as letters or graphemes. Therefore, graphemes are hypothesised to be relatively more important in shallow compared to deep orthographies.

---

[2] Ziegler and Goswami (2005) argue for a continuum in the size of the units involved in orthographic-phonological correspondences, with letter-phoneme correspondences on the smallest level, through body-rime correspondences, syllables, and whole words providing the largest units (see their Fig. 1). It is not always the case that the size of the units, as measured by the number of letters, is consistent with Ziegler and Goswami's (2005) terminology: for example, both the word *at* and the body *-at* are smaller than the grapheme *ough* (which is a grapheme, because it maps onto a single phoneme). We therefore talk about the "type" of units, rather than their size, and about grapheme-to-phoneme and body-rime correspondences, rather than small and large units.)

[3] A grapheme is a letter or letter cluster that maps onto a phoneme, or the smallest unit of speech. Examples of English graphemes are *t*, *th*, *ough*.

Although some studies support the PGST, open questions remain about the role of bodies in shallow orthographies, and to what extent the cross-linguistic differences in reliance on bodies is reliable. Evidence for the differential reliance on bodies as a function of orthographic depth comes from studies on the body-N effect: for a given letter string, body-N is the number of real words which have the same body (e.g., *at*, *hat*, and *brat* are all body neighbours of the word *cat*). The rationale behind body-N studies is as follows: if bodies are functional units underlying reading, then words with many body neighbours (e.g., *house*, which has a body neighbourhood of 6) should be easier to read than matched words with fewer body neighbours (e.g., *horse*, which only has one body neighbour).

Indeed, studies have shown a stronger body-N effect (faster reading latencies associated with high body-N) for English compared to German readers (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). This is in line with the PGST, because English is generally considered to be deep orthography, while the German correspondences are relatively unambiguous (Borgwaldt et al., 2005; Seymour et al., 2003; Ziegler et al., 2000).

Taking a close look at the results of the German participants only in the studies of Ziegler et al. (2001, 2003) shows a mixed picture. In the adult study (Ziegler et al., 2001), German readers showed little evidence for a body-N effect (the effect was marginally significant by subjects, and not significant by items). In developmental study, which used the same items (Ziegler, Perry, Ma-Wyatt, et al., 2003), there was a body-N effect for all groups of German readers: for dyslexic children, as well as the chronological-age and reading-age matched controls. Therefore, it is unclear whether adults show any reliance on bodies, or whether the study of Ziegler et al. (2001) did not have sufficient power to obtain a significant body-N effect for the German adults. Furthermore, recent evidence

suggests that body-N manipulations might not be particularly sensitive measures of reliance on bodies (see Paper 2). We therefore seek to address this issue with an alternative marker effect of large-unit processing.

In the current study, we aim to further explore the use of body-rime and grapheme-phoneme correspondences in English and German adult readers. In the current study, we use a body *consistency* manipulation to assess the degree of reliance on body-rime correspondences, and a *regularity* manipulation to assess the degree of reliance on grapheme-phoneme correspondence (GPC) rules.

On the linguistic level, the two concepts of body consistency and regularity are related and heavily confounded, but they are assumed to reflect different cognitive processes. Body consistency (hereafter: consistency) relates to the presence or absence of a word with the same body, but a different pronunciation. For example, the word *calm* is consistent, because all words with the body *-alm* are pronounced the same way; conversely, the word *warm* is inconsistent, because it has "enemies" with the same body but a different pronunciation, such as the word *harm*. Previous studies have found consistency effects, where words with inconsistent bodies (*harm*, *warm*) are read more slowly than matched words with consistent bodies (*calm*, *lung*) (Andrews, 1982; Jared, 1997, 2002; Jared, McRae, & Seidenberg, 1990).

In contrast, irregular words are words whose pronunciation does not comply to a set of GPC rules. GPC rules are generally defined using the sublexical rules of the Dual Route Cascaded (DRC) model (M. Coltheart et al., 2001), which implements each sublexical correspondence as the phoneme that most frequently co-occurs with a given grapheme. For example, the grapheme *a* is most often pronounced as "/æ/" (as in *cat*). Thus, the word *calm* would be considered irregular, because according to the rules, its pronunciation should be "/kælm/". GPCs also include multi-letter rules: for example, the

word "harm" does not contain the grapheme *a* because the letter *a* is part of a two-letter grapheme, *ar* [4]; as the grapheme *ar* is generally pronounced "/ɐː/", as in "car", the word *harm* is considered regular. Conversely, the word *warm* is irregular according to the GPCs, because it contains the same grapheme, *ar* and it is not pronounced as "/ɐː/" but as "/oː/". Thus, *calm* is irregular but consistent, *warm* is irregular and inconsistent; a word like *harm* is regular but inconsistent, and a word like *lung* is both regular and consistent.

Studies on the regularity effect compare reading latencies of regular words (*harm*, *lung*) to irregular words (*warm*, *calm*), and generally find slower reaction times for irregular compared to regular words in reading aloud (Andrews, 1982; Rastle & Coltheart, 1999; Roberts, Rastle, Coltheart, & Besner, 2003). The regularity effect has been explained within the DRC model (M. Coltheart et al., 2001), where reading involves two parallel procedures: The sublexical procedure operates on a set of GPC rules to decode a word in a serial manner, and is essential for the correct reading-aloud of nonwords. The lexical procedure retrieves a relevant entry for a given word from a mental lexicon, and is essential for the correct reading-aloud of irregular words because the sublexical decoding procedure will provide a regularised response, such as pronouncing *calm* as "/kælm/". The regularity effect is typically only found for words of low frequency, as the pronunciations of high-frequency words are retrieved quickly by the lexical route, before the misleading sublexical information can interfere (Metsala, Stanovich, & Brown, 1998).

As consistency relates to the reliability of the body, we use body consistency effects as a marker of body-rime processing. Regularity refers to the reliability of graphemes, and therefore we use it a measure of grapheme-phoneme processing. Previous research has shown that consistency and regularity independently affect reading speed

[4] This is according to Australian English; in rhotic dialects such as American English, *ar* would be represented as a context-sensitive rule, *a[r]* --> "/ɑ/".

(Andrews, 1982; Jared, 2002; Seidenberg, Waters, Barnes, & Tanenhaus, 1984). This indicates that different types of units are used by the sublexical route in parallel (cf. Schmalz et al., 2014; Paper 4).

Little research on regularity has been conducted in languages other than English, and the existing studies do not control for consistency as rigorously as English studies tend to[5]. The main reason for the small number if regularity studies across languages is the lack of irregular - and therefore the lack of inconsistent - words in most alphabetic orthographies. English, by definition of a deep orthography, contains many irregular words. In contrast, shallow orthographies such as German or Finnish have a very close correspondence between the written and the spoken word forms. A high number of irregular and inconsistent words in a given orthography are intrinsically linked for the following reason: the presence of inconsistent words in an orthography inevitably leads to the presence of irregular words, as only one body pronunciation can be considered regular. As a result, shallow orthographies differ from deep orthographies both in the number of irregular and inconsistent words.

Even though in shallow orthographies, irregularity or inconsistency occur relatively rarely, assessing their effect is of theoretical importance. Here, we aim to asses consistency and regularity effects as a function of orthographic depth, in order to assess reliance on large and small units respectively. Specifically, it is of interest whether readers of orthographies with a high degree of unpredictability due to irregular or inconsistent words develop quantitatively different reading mechanisms, such as increased reliance on lexical processes (Katz & Frost, 1992), as would be evidenced if we find a cross-linguistic difference in the size of the regularity effect, or qualitatively different mechanisms, as expressed by reliance on different types of sublexical units

[5] We know of no studies in orthographies other than English on the body-consistency effect.

(Ziegler & Goswami, 2005; Ziegler et al., 2001), as would be evidenced by a cross-linguistic difference in the size of the consistency effect, or both.

In the current study, we aim to compare the effects of consistency and regularity between two orthographies that are generally considered to vary in terms of depth, namely German and English (Borgwaldt et al., 2005; Seymour et al., 2003). We first conduct a direct comparison of the size of the consistency effect across English and German. We follow up with three experiments that attempt to establish to what extent the differences in the nature of irregularities across orthographies limit a cross-linguistic comparison of the regularity effect.

Two theories make explicit predictions about cognitive processes underlying reading as a function of orthographic depth. The first of these is the Orthographic Depth Hypothesis (ODH; Katz & Frost, 1992). This theory is based on a dual-route framework, where a letter string can be either pronounced by sublexical decoding, or by directly accessing whole-word (lexical) information. The sublexical output is, in most cases, sufficient for assembling the correct pronunciation of a word for shallow orthographies, while readers of deep orthographies are more dependent on the lexical procedure to get the pronunciations of irregular words, which do not comply with the sublexical rules (e.g., *yacht*, which would be pronounced as "/yætʃt/" if read by print-speech correspondences). Therefore, the ODH states that a greater ratio of lexical-to-sublexical processing is used for deep compared to shallow orthographies. In other words, the ODH proposes a quantitative difference in reading mechanisms across orthographies, where the ratio of lexical to sublexical processing varies as a continuum as a function of orthographic depth (Frost et al., 1987). As a result, we predict that readers of shallow orthographies should show a stronger regularity effect, because the lexical route does not provide the correct information for irregular words as quickly as it does in deep

orthographies and stronger interference of the misleading sublexical information should occur.

The second theory of cross-linguistic differences in reading is the PGST (Ziegler & Goswami, 2005). The PGST does not make any predictions about the relative use of the lexical versus the sublexical route; instead, it proposes qualitative differences in the nature of sublexical processing (Ziegler et al., 2001). Specifically, readers of deep orthographies are proposed to rely on larger sublexical units, such as body-rime correspondences, than readers of shallow orthographies. As such, we can make predictions, based on the PGST, about the size of the consistency effect: if readers of English rely more on body-rime correspondences, they should show a larger consistency effect than readers of shallow orthographies (e.g., German). In terms of the size of the regularity effect, if readers of shallow orthographies rely more on GPCs ("small" units) than readers of deep orthographies, we should find a stronger regularity effect in German than in English. The prediction regarding the size of the regularity effect as a function of orthographic depth is therefore identical according to both cross-linguistic theories, but only the PGST makes predictions about the size of the consistency effect across orthographies.

From previous studies, it is difficult to draw conclusions regarding the relative size of the regularity effect across languages. Firstly, there are several conflicting results from the few studies that exist of the regularity effect in languages other than English. Two studies have shown that, in contrast to English, there is no regularity-by-frequency interaction in French, a language which is considered to be of intermediate depth (Content, 1991; Ziegler, Perry, & Coltheart, 2003). Instead, irregular words of all frequencies are processed more slowly than regular words. This would indicate that there is considerable influence from the sublexical route, even for high-frequency words. Some

conflicting results have been found in a sample of Brazilian Portuguese-speaking children, who showed no regularity effect, either for low or high frequency words (Justi & Justi, 2009). Portuguese, like French, is considered to be a language of intermediate depth (Seymour et al., 2003; Sucena, Castro, & Seymour, 2009).

The second complication in comparing regularity effects across languages stems from cross-linguistic differences in the nature of irregularities. As an explanation for their lack of regularity effect, Justi and Justi (2009) suggest that their irregularities were too weak to affect word recognition. English contains strongly irregular words, such as *yacht* and *laugh*, while Portuguese irregular words tend to involve an ambiguity in the pronunciation of one letter only. Thus, the difference in the nature of the GPCs and irregularities across languages may not allow for a direct comparison of the regularity effect. [6]

In the current study, we address the issue of behavioural differences associated with the nature of the irregularities in two within-language experiments (3 & 4). Using a within-language comparison, we aim to establish the degree to which the nature of irregularity matters for single-word reading aloud. As we are reluctant to perform a direct cross-linguistic comparison between English and German, we present a German regularity manipulation in a separate experiment (2). This allows us to independently assess the size and characteristics of the regularity effect within a language, and draw some general conclusions about the differences and similarities across languages based on the patterns of results.

Before moving on to describing the experiments, we provide an overview the nature of German irregularities (for a more detailed description, see Ziegler et al., 2000).

[6] This problem does not apply for consistency effects. As long as the orthography contains any number of inconsistent words, consistency can be quantified by the ratio of friends to enemies, which is not dependent on the characteristics of the orthographic system.

Two major classes of irregularities are *loanwords*, and *subtle irregularities*. Loanwords are mainly derived from French and English (e.g., *Charme, Jazz*). They mostly conform to the GPCs of their language of origin, but not to the German GPCs. These irregular words cannot be used to study regularity effects in German undergraduate students, as this population generally has at least some knowledge of English and French and the participants would be more likely to switch to another set of GPC rules rather than read the items in a way that an English reader would read an item such as *laugh* (cf. Treiman, Kessler, & Evans, 2007).

The second major class of irregular words in German have been termed subtle irregularities (Ziegler et al., 2000). Subtle irregularities are violations of a specific set of GPC rules referred to as super-rules. A super-rule differs from other GPCs because it applies to a group of letters with a given characteristic (e.g., vowels), as opposed to a specific grapheme. Of interest here are two rules which can be used to determine vowel length: in a monosyllabic word, when a vowel is followed by only one consonant, it is pronounced as a long vowel (e.g., *Wal*, "/vaːl/", whale), and if it is followed by two or more consonants, it is pronounced as a short vowel (e.g., *Wald*, "/valt/", forest). There are exceptions to this rule, such as the words *Bus* (“/bʊs/”, same meaning as in English) and *Keks* ("/keːks/", cookie). These words that break the super-rules are called subtly irregular words.

Subtle irregularities differ from English irregular words for several reasons. Firstly, there are no comparable super-rules in English, and it is unclear how super-rules are processed by the cognitive system - or, in fact, whether there is any psychological reality to this type of GPC rule (Perry, Ziegler, Braun, et al., 2010). Secondly, the irregularity relates to a difference in a single phoneme, namely vowel length. Given English irregular words like *yacht*, where the discrepancy between the regularised

("/jætʃt/") and correct pronunciation ("/jɔt/") is not at all subtle, it is intuitive to expect that the English readers would show a larger regularity effect, compared to German readers for these subtle irregularities associated with vowel length. However, there are also some English words that are more weakly irregular than the word *yacht*. The irregular word *lounge*, for example, only has an irregular pronunciation of the final grapheme (it is pronounced as "/ʒ/", when the rule-based pronunciation would be "/dʒ/").

In summary, given the state of knowledge, as described above, several questions remain. Firstly, it is unclear whether German readers rely on body units at all. Even when it comes to English, there are some mixed results from previous studies (see Paper 2). If there is any psychological reality to the concept of body units, we should find a main effect of body consistency. If reliance on bodies is greater in English than German, we should furthermore find an interaction between language and body consistency, with a stronger effect for English than German. In fact, we know of no study that has shown any evidence for reliance on bodies in German - even the study of Ziegler et al. (2001) found no reliable body-N effect in a follow-up test using the German items only. Therefore, it is unclear whether we will find a consistency effect in German at all. We examine these questions in Experiment 1.

In terms of the regularity effect, it is unclear how its size differs across orthographies varying in depth, as the previous literature provides mixed results. We also know very little about the ways in which different characteristics of irregularities influence reading aloud, as regularity is generally defined as a binary distinction, where a word is either regular (i.e., it complies to the rules) or irregular (i.e., it does not; M. Coltheart, 2012). Before conducting any direct cross-linguistic comparisons of the size of the regularity effect, one needs to establish whether the nature of irregularities across orthographies is comparable. One intuitively appealing possibility is that the overlap

between the regularised and correct version of a word would affect reading latencies; for example, it might be easier to derive the correct pronunciation of a word like *lounge* (where the sublexical information predicts the regularised pronunciation "/læɔndʒ/", and the correct output is "/læɔnʒ/") compared to a word like *laugh* (where the GPCs give the regularised pronunciation "/loː/", while the correct pronunciation is "/lɐːf/"). In Experiment 2, we seek to establish whether we are able to produce a regularity effect in a shallow orthography, namely German. In Experiments 3 and 4, we follow up with within-language investigations in English, assessing to what extent the nature of the irregularities, as measured by the degree of overlap between the regularised and correct pronunciation, affect reading latencies.

1.2. Experiment 1: Consistency effects in German and English

The aim of the first experiment was to use the consistency effect as a marker of reliance on body units in German and English. We aimed to test (1) whether we will find a consistency effect in German, given some mixed results from previous studies on the body-N effect, and (2) whether we find a stronger consistency effect for English than German, as predicted by the PGST (Ziegler & Goswami, 2005).

1.2.1. Methods

The participants were 23 English native speakers who were undergraduate students at Macquarie University, and 16 German native speakers recruited via a snowball method. All German participants had completed their primary and secondary schooling in Germany. Four of the German participants were in Australia for a short-term visit, the others were residents of Australia. Each participant was tested in their native language.

For the item set, we only used regular words, where the pronunciations corresponded with the GPCs of the English and German DRCs respectively (M. Coltheart

et al., 2001; Ziegler et al., 2000). Consistent words were defined as having no enemies, or words with the same body but a different pronunciation. Inconsistent words were items which had at least as many enemies as friends, i.e., a type consistency ratio of 0.5 or smaller (as determined by corpus analyses conducted by Ziegler, Stone, & Jacobs, 1997 for English, and J. Ziegler, personal communication, 2012, for German). The resulting items were 17 pairs of words in each language. These are listed in the Appendix; the item descriptive statistics can be found in Table 1.

Table 1.

Descriptive statistics of the consistent and inconsistent conditions in English and German; mean (SD).

	German		English	
	Consistent	Inconsistent	Consistent	Inconsistent
Number of letters	4.53 (1.37)	4.53 (1.37)	4.29 (0.85)	4.29 (0.85)
Orthographic N	4.94 (4.92)	3.47 (2.98)	6.76 (4.24)	7.06 (5.36)
Number of friends	6.29 (4.70)	2.12 (1.17)	8.29 (5.65)	2.59 (1.18)
Number of enemies	0 (0)	2.18 (1.29)	0 (0)	4.35 (2.89)
CELEX frequency	69.29 (106.11)	49.12 (60.52)	271.06 (655.22)	260.88 (615.12)
Token consistency	1 (0)	0.48 (0.36)	1 (0)	0.27 (0.19)

The items were presented in random order with the program DMDX (Forster & Forster, 2003) in uppercase letters, in light grey on a dark background. A trial consisted of a fixation cross (700ms), followed by the item, which stayed on the screen for 1500 ms

or until the voice-key was triggered. The participants were instructed to read aloud each word as quickly and accurately as possible. Ten practice items preceded the experiment.

1.2.2. Results

The responses were marked off-line with the program CheckVocal (Protopapas, 2007) as correct, incorrect, or non-response. We excluded all non-responses (3.38%). Before conducting the RT analyses, we removed all incorrect responses (3.16%) and all data points which deviated more than 2 SDs from each participant's mean (5.34%). Table 2 shows the RTs and accuracy for German and English across the two conditions.

We used Linear Mixed Effect (LME) models (Baayen, 2008; Baayen, Davidson, & Bates, 2008) to assess the effects of language and consistency, and their interactions. We analysed the error rate and latencies of the English and German consistent and inconsistent words, yielding a 2-by-2 factorial design.

Table 2.

Reaction times and accuracies of the German and English consistent and inconsistent words; mean (SD).

	German		English	
	Consistent	Inconsistent	Consistent	Inconsistent
RT (ms)	530.9 (98.7)	533.9 (103.5)	498.5 (131.2)	489.6 (126.7)
Accuracy (%)	99.63 (6.06)	95.96 (19.74)	99.23 (8.77)	92.67 (26.10)

The accuracy analysis showed a significant effect of consistency, $z = -2.1$, $p < 0.04$. Neither the effect of language, nor the language-by-condition interaction approached significance, $p > 0.6$. As can be seen from Table 2, inconsistent words were read aloud less accurately than consistent words in both English and German. For the RT analyses, neither the effects of condition or language, nor their interaction approached significance, all $p > 0.4$.

Especially given the small number of items and participants, it is impossible to draw conclusions from the non-significant interaction between language and consistency: it could either reflect an effect of the same magnitude in the two orthographies, or insufficient power to detect an interaction. We therefore followed up with a Bayes Factor to assess the strength of evidence against this theoretically important interaction (Morey & Rouder, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009). We compared the evidence for a full model, which contained the two main effects and their interaction, to a model including the main effects only. In accuracy, the Bayes Factor provided evidence against the full model and thus against the presence of an interaction, with a Bayes Factor value of 0.24 (±2.74%), where a value smaller than 0.3 is considered to provide evidence against the model that is being tested (Rouder et al., 2009). In RT, the Bayes Factor value for the interaction was nearly identical to the accuracy value: 0.22 (±4.69%).

1.2.3. Discussion

In Experiment 1, we found that consistency affects naming accuracy, but not naming speed. Similarly, in a previous consistency experiment both consistency and regularity effects were stronger in accuracy than reaction times in a reading aloud task (Jared, 2002). A consistency effect in accuracy but not in reaction times may occur when the items have a relatively low frequency. If entries in the orthographic lexicon are not well established, the output from the sublexical route may occur before the lexical look-up procedure is complete. If bodies are used as a sublexical mechanism to predict the pronunciation of an item, the incorrect response is given for words with more enemies than friends before the orthographic lexicon can correct it.

We found a non-significant interaction between consistency and language. A Bayes Factor analysis supported the view that there is no interaction. Overall, this shows the utility of Bayes Factor analyses in psycholinguistic research: due to the small number

of items (which was unavoidable, due to the small number of inconsistent words in German) it would otherwise have been impossible to assess the possibility that there is no interaction. The lack of an interaction has theoretical implications: it suggests that bodies are equally salient in both English and German, which questions the assertion that reliance on bodies differs as a function of orthographic depth (Ziegler & Goswami, 2005). This issue is taken up in more detail in other parts of this thesis (see Paper 2 and 3), so we do not discuss it further here.

The data of Experiment 1 confirms the presence of a consistency effect in German: as we controlled for GPC regularity, this shows that there is reliance on body-rime correspondences even in an orthography with relatively unambiguous print-to-speech correspondences. Evidence for reliance on bodies in German using a consistency manipulation raises two questions: namely, (1) why we find reliance on bodies in a shallow orthography, such as German, and (2) how our lack of consistency by language interaction can be reconciled with previous research, which showed a body-N by language interaction (Ziegler et al., 2001). We discuss these in turn below.

Firstly, it is unclear why readers of German would show reliance on body-rime correspondences, given that GPCs provide the correct output for most words. According to the PGST (Ziegler & Goswami, 2005), reliance on larger units develops due to the ambiguity of small unit correspondences. Goswami and Ziegler (2006) acknowledge that even in shallow orthographies, reliance on large units may develop, but they are referring to whole-word units: in contrast to sublexical units, whole words give the reader direct access to meaning, and are therefore they are useful for readers of all orthographies. This reasoning does not apply to large sublexical units: the body unit *-ung*, for example, does not contain any more semantic information than the grapheme *ng*.

LaBerge and Samuels (1974), in a theory of automaticity in reading, propose that reliance on larger sublexical clusters, such as bodies or syllables, develops in order to increase reading fluency: if a reader learns to decode a body (e.g., *-ung*) with equal speed to a single letter or grapheme (e.g., *u* or *ng*), the sublexical reading process will speed up; the recognition of the sublexical cluster *-ung* will be two times faster if body-rime correspondences are used compared to the two individual graphemes. Several authors have pointed out that the increase in reading fluency across reading development is not considered by the current version of the PGST (de Jong, 2006; Wimmer, 2006). Developmental studies are needed to study the progression of smaller to larger units during reading acquisition and how this differs across orthographies - as the current study focuses on skilled reading, we cannot address this question. As a working hypothesis, we propose that both the need to reduce the unpredictability associated with the pronunciation of an unfamiliar word and the need to increase fluency provide two independent pressures on a child learning to read to rely on units of different types, and in particular units that contain more letters than the letter-to-sound correspondences that are explicitly taught at the beginning of reading instruction.

The second question is why we found evidence for reliance on body units in German when we used a body consistency manipulation, compared to the previous studies using a body-N effect (Ziegler et al., 2001; see also Paper 2). A theoretically interesting possibility is that the body-N effect and the consistency effect measure slightly different constructs. The body-N effect may directly relate to the saliency of the orthographic units: it is an approximate measure of exposure to the particular orthographic cluster. Consistency does not take into account this frequency information: rather, it specifically taps into the connection between the orthographic and phonological unit. If the connection is unequivocal (i.e., consistent), it is likely to be stronger compared

to a situation where an orthographic cluster has two potential phonemic representations (i.e., for an inconsistent body). Thus, it is possible that Ziegler et al., (2001) showed increased saliency of orthographic body units in English compared to German, rather than a difference in the strength of connections between bodies and rimes. This may not even be purely because of the ambiguity of letter-to-sound correspondences: bodies in English have an overall higher frequency because the English orthography is denser (i.e., English words, on average, have more letter-substitution neighbours than German words; see Baayen, Piepenbrock, & Gulikers, 1995). This may affect the speed at which a particular orthographic unit is activated, but not necessarily the strength of the connection between an orthographic and a phonological unit. Such a possibility would need to be followed up by future research with more sophisticated methodology.

It is worth noting that previous studies have found stronger consistency effects when manipulating the consistency ratio by token rather than by type (Jared, 1997, 2002; Jared et al., 1990). This way of calculating the consistency ratio takes into account the frequency of the item's friends (body neighbours with the same pronunciation) and enemies (body neighbours with a different pronunciation), rather than the ratio of the number of friends and enemies. Due to the restricted number of items, it was impossible to match the items across languages on the token consistency ratio. As a result, the German inconsistent items have a higher token consistency ratio than the English inconsistent items. This would predict a smaller consistency effect in German readers; but this was not what we found. While it may be suggested that the difference in token consistency could have masked a larger consistency effect in German than in English,[7] we consider this unlikely, as there is no theoretical reason to expect such a difference.

[7] We thank an anonymous reviewer from a previous journal submission for pointing out this possibility.

In summary, we found a consistency effect in both English and German (in accuracy), suggesting reliance on larger grain sizes, or body-rime correspondences, in both orthographies. The next question we address is whether the regularity effect, as a measure of reliance on GPCs, behaves similarly across languages. We explore this question in two separate experiments for German (Experiment 2) and English (Experiments 3 and 4), because of the differences in the nature of irregularities across languages.

1.3. Experiment 2: Regularity effects in German

In Experiment 2, we aimed to examine whether we can find a regularity effect for German. German irregular words are either loanwords (e.g., *couch*) or subtle irregularities, where the vowel length is not predictable based on a set of super-rules (*V[C]* --> "long vowel", *V[C][C]* --> "short vowel"). The nature of this rule, and therefore the irregularity that is associated with non-compliance to this rule, is different from scenarios which have been well-studied in English. Therefore, it is unclear whether we will find a regularity effect for these subtly irregular words.

1.3.1. Methods

Eighteen German native speakers took part in Experiment 2: the task was reading aloud regular or irregular German words. As in Experiment 1, they were recruited by a snowballing method. All participants resided in Australia, but had attended school in Germany and had thus received some of their reading instruction in German. On average, they had attended a German school for 10.95 years (SD = 2.92).

Irregular words contained subtle irregularities, as described by Ziegler et al. (2000). We only used words with consistent bodies (J. Ziegler, personal communication). The final stimuli consisted of 18 words in each of the two conditions and are listed in the Appendix. The descriptive statistics for these items are shown in Table 3. They were

matched to each other on initial phoneme, letter length and frequency (Duyck, Desmet, Verbeke, & Brysbaert, 2004). The testing and scoring procedure was identical to Experiment 1.

Table 3.

Descriptive statistics of the regular and subtly irregular conditions in German; mean (SD).

	Regular	Irregular
Number of letters	4.06 (0.83)	3.72 (0.67)
Orthographic N	3.35 (2.52)	2.28 (2.61)
CELEX frequency	126.82 (296.52)	158.22 (418.14)

1.3.2. Results

The RTs and accuracy for the experiment are shown in Table 4. For each inaccurate item, we also recorded the type of error. One trial was excluded as a non-response, and 67 responses were incorrect (10.36%). Of particular interest were errors relating to vowel length, as these represent the regularisation errors in the subtly irregular condition.

Table 4.

Reaction times and accuracy for the regular and subtly irregular conditions in German; mean (SD).

	Regular	Irregular
RT (ms)	549.6 (140.8)	567.7 (125.9)
Accuracy (%)	93.17 (25.27)	88.99 (11.01)

Errors in vowel length made up 78% of all incorrect responses. In the subtly irregular condition, 79% of all errors related to vowel length, compared to 75% in the regular condition. This difference was non-significant, $t(34) = 1.00$, $p > 0.3$, suggesting that vowel length errors occur relatively often in German compared to other types of errors, and do not depend on compliance to the super-rules. To explore whether there were any overall differences between the regular and subtly irregular conditions, we

conducted an LME. The regularity effect was not significant in accuracy, $z < 1$ or RTs, $t < 1$.

1.3.3. Discussion

As outlined in the introduction, it was unclear whether such an effect exists, because the nature of subtle irregularities, and the rules that underlie these, is different from the irregular words of English. Indeed, we found no regularity effect in German, which suggests that subtle irregularities are processed differently from irregular words in English (cf. Perry, Ziegler, Braun, et al., 2010; Schmalz et al., 2014. i.e., Paper 4).

Aside from the obvious limitation of our small item set, [8] there are two possible explanations for the null-result of Experiment 2. Firstly, it is unclear whether there is any psychological reality to the super-rules. Perry et al. (2010) compared two computational models, namely the German DRC, which contained the super-rules, and the German CDP+, which derives the sublexical correspondences using a learning algorithm, to nonword responses of German undergraduate students. The CDP+ performed better at predicting the vowel length responses of participants than the DRC, suggesting, at least, that super-rules are not the only determiner of vowel length pronunciation in German. Schmalz et al. (2014; Paper 4) used another nonword reading task and found that German participants relied on super-rules to some extent, but they also relied on single-letter rules (with an overall bias to short rather than long vowels) and body-rime correspondences. If super-rules are, at the very least, not the only sublexical correspondences that are used by skilled readers, we should not necessarily expect that words which do not comply with them are harder to read than words that do.

[8] This was unavoidable, because of a limited number of irregular words in the German language and our care in matching for potential confounds.

A second possibility is that the difference between the regularised and correct pronunciation of a subtly irregular word may be so small that the conflict between the lexical and sublexical route does not have a noticeable effect of reading aloud performance (Justi & Justi, 2009). German subtly irregular words contain only one irregular correspondence, and there are only two plausible ways of pronouncing the ambiguous letter. Therefore, it is possible that the discrepancy between the regularised and correct pronunciation is not large enough to yield a conflict between the lexical and sublexical route.

1.4. Experiment 3A: Regularity effects in English

In Experiment 2, we concluded that German super-rules (and subtle irregularities) are different in nature compared to English GPCs (and English irregular words), but the German experiment does not allow us to establish what it is specifically why this cross-linguistic difference occurs. One possibility that has been suggested by Justi and Justi (2009) is that a regularity may only be found when the difference between the regularised and correct pronunciation of an irregular word is substantial.

A way to test whether this explanation for our German null-result is viable is by conducting an equivalent experiment in English: due to the larger number of irregular words and the variability of their nature, we can manipulate the degree of deviation of the correct to the regularised pronunciation. English contains some strongly irregular words, such as the words *yacht* or *laugh*, which hardly have any relationship between the spoken and written form, but also weaker irregularities such as *lounge* or *learn* (where the second grapheme, *ear*, has an irregular pronunciation). If the lack of a regularity effect in German is due to the weak nature of the irregular words, we expect that in English, a regularity effect would only be found for strongly irregular words, but not for weakly irregular words.

1.4.1. Methods

The participants were 22 native English speakers who were staff or students at Macquarie University. There were three conditions of stimuli: regular, weakly irregular and strongly irregular words. Regular words were defined as complying with the GPC rules of the English DRC (M. Coltheart et al., 2001). In order to objectively classify the irregular words as weakly or strongly irregular, we used the DRC (Coltheart, et al., 2001) to obtain both the regularised pronunciation, and the correct pronunciation for the irregular words. The degree of regularity of each word was then calculated as the number of phonemes where the correct pronunciation deviates from the regularised version, divided by the total number of phonemes of the correct pronunciation. Strongly irregular words were defined as having a regularity proportion of 0.5 or less, and as weakly irregular if the regularity measure was greater than 0.5. [9]

As in Experiment 2, we used only consistent words (Ziegler et al., 1997). The resulting stimuli were 3 conditions of 18 words each. The item set of Experiment 3A is small, because we originally matched the regular/irregular items across languages, which strongly limited the selection process. The items are listed in the Appendix and the descriptive statistics in Table 5. The testing and scoring procedures were identical to the previous experiments.

1.4.2. Results

Fifteen trials (1.2% of the data) were excluded due to poor sound quality. Table 6 shows the accuracy and RTs for each of the three conditions (along with the results of Experiment 3B). Overall, there were 50 errors (4.1% of all data), 36 of which were regularisations. A significantly higher proportion of regularisation errors was made for

[9] Using the same procedure on the German items used in Experiment 2 showed that all items were indeed classified as weakly irregular.

strongly irregular than weakly irregular words, $t(34) = 2.07$, $p < 0.05$. An LME was used explore the differences between the regular, weakly irregular and strongly irregular conditions in English. The accuracy analysis showed that the effect of the strongly irregular condition compared to the regular condition was significant, $t = 2.49$, $p < 0.02$. The weakly irregular words, however, did not differ significantly from the regular condition, $p > 0.8$.

Table 5.

Descriptive statistics of the regular, weakly irregular, and strongly irregular conditions in English used for Experiment 3A; mean (SD).

	Regular	Weakly irregular	Strongly irregular
Number of letters	4.29 (0.59)	4.28 (0.75)	4.44 (0.70)
Orthographic N	6.18 (5.52)	7.11 (5.04)	3.17 (2.90)
CELEX frequency	295.06 (952.3)	192.34 (247.4)	66.94 (120.8)

Table 6.

Reaction times and accuracy of the regular, weakly irregular, and strongly irregular conditions in English obtained in Experiments 3A and 3B; mean (SD).

		Regular	Weakly irregular	Strongly irregular
3A	RT (ms)	474.8 (87.1)	469.1 (78.8)	397.0 (85.3)
	Accuracy (%)	98.99 (10.01)	99.47 (7.25)	90.26 (29.69)
3B	RT (ms)	504.8 (39.0)	539.5 (40.9)	542.9 (57.3)
	Accuracy (%)	99.20 (1.57)	92.26 (13.70)	84.04 (26.07)

The LME on RTs closely mirrored the accuracy analysis: latencies in the strongly irregular condition were significantly slower than latencies for the regular words, $t = 2.30$,

$p < 0.03$, while the weakly irregular condition was not significantly different from the regular one, $p > 0.8$.

1.4.3. Discussion

Experiment 3A served partly as a follow-up to Experiment 2: as it was unclear whether the lack of a regularity effect in German was due to the subtle nature of the irregularities, we manipulated the proportion of overlap between the regularised and correct pronunciation in English. Specifically, we tested whether it is the case that irregular words with a small discrepancy between the correct and regularized pronunciations are processed differently from irregular words for which this discrepancy is large. We found that regular words were read aloud more slowly and less accurately than strongly irregular words but that weakly irregular words were read with the same speed and accuracy as regular words.

1.5. Experiment 3B: Better controlled regularity study in English

From Experiment 3A, it looks like the degree to which the regularised and correct pronunciations overlap affects reading latencies. Before drawing any conclusions from these results, however, it is worth expanding upon a potential confound in the item set. Regularity was defined as deviation of the correct pronunciation to that given by the sublexical route of the DRC (M. Coltheart et al., 2001). Yet, behavioural evidence suggests that English readers rely to a greater extent on the context of each GPC than does the DRC (Schmalz et al., 2014, i.e., Paper 4; Treiman, Kessler, & Bick, 2003). Thus, for example, the word *learn*, one of the weakly irregular items from Experiment 3A, is irregular by DRC's GPC rules, but it is possible that participants do not process it as an irregular word, because it complies to the context-sensitive rule that *ear* is pronounced as

"/ɜː/" when followed by a consonant.[10] Such complex rules could resolve the pronunciation of numerous items in our weakly irregular condition, but not strongly irregular condition. We therefore conducted a further study, with a larger and more controlled item set.

1.5.1. Methods

As in Experiment 3A, there were three conditions of stimuli: regular, weakly irregular and strongly irregular words. We took care to include only irregular items where the pronunciation could not be resolved via context-sensitive correspondences[11]. Across condition, we matched for frequency, orthographic N, length, the number of neighbours that have a higher frequency (HFN), token body-rime consistency, and the position of the first irregularity within the irregular words (an important variable: see Rastle & Coltheart, 1999; Roberts et al., 2003). We also improved on the methodology of the previous experiment by ensuring that the words were familiar to the participants, as we used the English Lexicon Project database (Balota et al., 2007), to exclude words where participants made more than 10% lexical decision errors. Thus, all the words we used were likely to be known to our participants, which reduced the possibility that they were regularised because the correct pronunciation was not represented in their mental lexicons. The resulting items are listed in the Appendix, and the descriptive statistics are in Table 7.

[10] We thank Marcus Taft for pointing out this possibility.

[11] As it is still unclear what rules are used by the sublexical system, how individuals decide which rules to apply under what conditions, and how and why this varies across individuals (Pritchard, Coltheart, Palethorpe, & Castles, 2012; Schmalz et al., 2014), there is no systematic way of determining whether a word's pronunciation is predictable based on complex rules or not, especially in the English orthography. Therefore, this was decided based on the author's intuitions.

Table 7.

Descriptive statistics of the regular, weakly irregular, and strongly irregular conditions in English used for Experiment 3B; mean (SD).

	Regular	Weakly irregular	Strongly irregular
Number of letters	4.82 (0.72)	4.61 (0.96)	4.75 (0.84)
Orthographic N	4.57 (3.50)	3.11 (3.34)	3.68 (3.46)
CELEX frequency	235.9 (189.8)	215.3 (187.2)	204.7 (144.3)
Token consistency	0.84 (0.30)	0.73 (0.33)	0.78 (0.39)
Number of higher frequency neighbours	2.07 (1.84)	1.98 (2.30)	1.50 (2.30)

The participants were 27 undergraduate students at Macquarie University who participated in exchange for course credit. The presentation and scoring methods were identical to the previous experiments.

1.5.2. Results

Before conducting the RT analyses, we excluded all incorrect responses (7.23%) and data points which deviated more than 2 SD from the mean (3.40%). The average RTs and error rates for the three conditions are summarised in Table 6. The critical comparisons are (1) regular versus irregular words, to confirm that we find a regularity effect, and (2) weakly versus strongly irregular words, to explore whether there is any difference between the two conditions.

The comparison of the regular to the two irregular conditions showed a significant effect for both: in accuracy, both the strongly, $z = 3.56$, $p < 0.001$, and weakly, $z = 2.55$, $p < 0.02$, irregular words were read less accurately than words in the regular condition. This was mirrored in the RT analyses, where the strongly, $t = 3.00$, $p < 0.01$, and weakly, $t = 3.23$, $p < 0.01$, irregular words were read more slowly than regular words. Next we

compared the strongly versus weakly irregular words. Neither the accuracy nor the RT analyses showed a significant difference between these two conditions, $p > 0.8$.

1.5.3. Discussion

In Experiment 2 we showed no regularity effect in German, where the irregular words contained so-called "subtle irregularities" (Ziegler et al., 2000). In Experiment 3A we followed up with an English experiment to see whether the results of Experiment 2 could be explained by an insufficient deviation of the regularised to the correct pronunciation. Initially, this seemed to be the case, as in Experiment 3A, weakly irregular words with a small discrepancy between the regularised and correct pronunciations were not read aloud any more slowly than regular words by English readers, but strongly irregular words were read both more slowly and less accurately. Yet, Experiment 3B shows that once we include only irregular words where the pronunciation cannot be resolved by the use of context-sensitive rules, this difference between the two irregular condition disappears: both weakly and strongly irregular words are read aloud more slowly than regular words.

Though this departs somewhat from our original aims of conducting this study, Experiments 3A and 3B provide some information regarding how processing by the sublexical route operates. It seems that English words containing context-sensitive correspondences are not processed in the same way as irregular words. This has implications for defining irregularity: if an irregular word is defined as one where the pronunciation departs from the sublexical output of the DRC, it is clear that the GPC rules of the DRC should be upgraded to contain more context-sensitive rules (M. Coltheart, 2012; M. Coltheart et al., 2001; Schmalz et al., 2014). Future research is needed to establish how such context-sensitive rules are used by readers, and how computational models such as the DRC could be modified to reflect the cognitive

processes that underlie the decision making associated with assigning the correct pronunciation to words containing complex correspondences.

Our results also suggest that the degree of irregularity does not make a difference for reading aloud latencies or accuracy. This is in line with the DRC framework, and in particular previous studies on the position of regularity effect (M. Coltheart & Rastle, 1994; Havelka & Rastle, 2005; Rastle & Coltheart, 1999; Roberts et al., 2003). As the position of the first irregular correspondence was matched across the irregular conditions of Experiment 3B, a serial sublexical decoding process would be impaired at the same point in time when processing strongly and weakly irregular words. Strongly irregular words contain more than one irregular correspondence. Before the later-positioned irregular correspondences can interfere with phonological output processes, it is likely that the lexical route has already provided the correct pronunciation, so the misleading sublexical information from the irregular correspondences that occur in the later positions of a strongly irregular word would not interfere with naming. Such an explanation assumes serial processing of the sublexical route. In a model where sublexical correspondences are processed in parallel, such as the computational implementations of the triangle model (Harm & Seidenberg, 1999; Plaut, 1999; Plaut et al., 1996; Seidenberg & McClelland, 1989), both early and late occurring irregular correspondences should be processed simultaneously, so the degree of irregularity (number of irregular correspondences) should have a noticeable impact on reading performance.

1.6. General discussion

In the current study, we conducted four experiments, which had the following broad aims: (1) to further examine the reliance on body-rime and grapheme-to-phoneme sublexical correspondences in English and German, and (2) to specify the mechanisms

underlying the regularity effect. Although our results are not as clear-cut as we may have hoped, some conclusions can be drawn.

Firstly, we showed that skilled German readers rely on body-rime correspondences to some extent, because, like the English readers, they showed body consistency effects in accuracy. We propose that even in an orthography with relatively unambiguous print-to-speech correspondences, reliance on correspondences that other than GPCs develops in order to increase reading fluency. Although similar claims have been made for over a century (e.g., Huey, 1908; LaBerge & Samuels, 1974; Prinzmetal, Treiman, & Rho, 1986), this possibility has not been addressed by more recent theoretical work, such as the psycholinguistic grain size theory (Ziegler & Goswami, 2005).

Secondly, we attempted to examine the regularity effect in both German and English. Though it may be expected that the regularity effect is stronger in English than in German, we were unable to conduct a direct cross-linguistic comparison due to the different nature of the GPC rules in the two languages. Nevertheless, we found no regularity effect in German when we used words that did not comply with Ziegler et al.'s (2000) super-rules as irregular words. Most likely, this is because there is limited psychological reality to the super-rules. [12]

The follow-up studies that we conducted in English raise some issues and questions. From a methodological perspective, we show that in an irregularity manipulation, it is important to exclude words where the pronunciation can be resolved by relying on context-sensitive rules. A logical follow-up study might aim to establish, in a more controlled environment, how the presence of context-sensitive rules affects

[12] One of the original aims of this study was to establish whether super-rules have any psychological reality. In that sense, the current experiment is not a complete failure, although a different paradigm, such as the computational study of Perry et al. (2010), or the nonword reading study and optimisation procedure by Schmalz et al. (2014), provides much clearer evidence.

reading aloud both compared to irregular words with unpredictable pronunciations and compared to regular words with simple GPCs only. However, this is outside the scope of this thesis. Practically speaking, this would be very difficult if not impossible in English, because the complexity of rules and their unpredictability is often difficult to dissociate.

In summary, future research needs to: (1) explain why German readers, like English readers, rely on body-rime correspondences, even though in German GPCs would be largely sufficient for accurate reading, (2) take into account differences in the structures of orthographic systems which often limit the conclusions that can be drawn from cross-linguistic studies, and (3) make a distinction between "irregular" words that can be resolved via more complex correspondences, and those where the pronunciation is not predictable based on any type of sublexical correspondence.

Appendix: Items used in the experiments

Experiment 1

English consistent

board, boat, feet, fraud, fuss, gang, hope, kneel, life, lump, morse, my, myth, quit, reed, scale, wright

English inconsistent

bead, beard, cost, food, foul, frost, golf, here, knead, limb, lost, mood, no, nonce, read, scarf, wreath

German consistent

bös, gar, grau, hell, Leu, Lot, Macht, Maß, mies, Punkt, Schal, Schirm, schlimm, Schrank, Schrei, Spuk

German inconsistent

blond, Bruch, Bub, Gen, grob, Herz, Lob, los, manch, Mus, Nerz, Scherz, Schmerz, Schub, Schwarz, Spruch, stet

Experiment 2

Regular

Amt, Brom, bunt, Burg, elf, falsch, halb, jetzt, jüngst, kalt, Kern, Kinn, Lid, Lurch, Skalp, Tod, trüb

Irregular

Arzt, bis, Bit, Box, erst, flugs, Herd, hin, Jagd, Keks, Koks, Krebs, Lok, Lux, Slip, tags, Tip, Yard

Experiment 3A

Regular

angst, arch, bag, bark, broom, coach, coast, fact, herb, herd, kiss, lens, lift, scalp, term, toast, yawn, yes

Weakly irregular

ask, axe, bask, bolt, brief, cast, cold, cook, field, hearse, hook, last, learn, salt, task, told, yearn, young

Strongly irregular

aisle, balm, beau, beige, calf, calm, corps, folk, half, heir, her, laugh, lisle, psalm, talk, tongue, yacht, yeah

Experiment 3B

Regular

airs, bait, bombed, flaws, flint, grade, grant, hoarse, hoops, hull, hushed, jets, knee, melt, noon, plus, skirt, squeal, stamp, starts, storm, stud, swore, tiles, twist, weeds, wheat, wrap

Weakly irregular

aunt, axe, chef, clerk, comb, deaf, dealt, debt, dove, gauge, gear, grind, gym, height, monk, niche, plague, plaque, priest, roll, seize, shriek, ski, sword, thief, tomb, warmth, wolf

Strongly irregular

ache, aisle, calf, chalk, corps, cough, dough, earn, ease, folk, laughs, mayor, palm, prayer, rouge, sew, shoe, sioux, stalk, suave, sue, suede, suite, theirs, tongue, truths, vase, weird.

Paper 2: A meta-analysis of body-N effects

Abstract

Previous research has indicated that words and nonwords with many body neighbours (i.e., words with the same body, e.g., *cat, brat, at*) are processed faster than words with few body neighbours. This is known as the body-N effect, and has been previously shown in lexical decision latencies (Ziegler & Perry, 1998) and in reading aloud (Ziegler et al., 2001). Our aim was to provide a thorough analysis of the body-N effect by conducting a meta-analysis of the data from 11 item sets from different studies. Using Linear Mixed Effect Model and Bayes Factor analyses, we show that an inhibitory body-N effect emerges for nonwords in a lexical decision task, but that body-N has no influence on lexical decisions for words, or on the reading aloud of words or nonwords. The results and patterns across languages, in German and English, were strikingly similar. The findings have implications for models of reading, as they suggest that the locus of the effect is lexical and could reflect a hierarchical structure of the orthographic lexicon. The null-results for the reading aloud tasks, and the lack of any cross-linguistic differences, have implications for a cross-linguistic theory of reading, the Psycholinguistic Grain Size Theory (Ziegler & Goswami, 2005), as its key assumption of stronger reliance on bodies in English than German readers has been provided using this body-N manipulation.

2.1. A meta-analysis of body-N effects

It is well established that the processing of a given word or nonword is influenced by the existence of similar-looking words (Andrews, 1997; M. Coltheart, Davelaar, Jonasson, & Besner, 1977; Yarkoni, Balota, & Yap, 2008; Ziegler & Perry, 1998). Different metrics have been used to capture the degree to which a letter string resembles other words. The original concept was introduced as orthographic neighbourhood (Coltheart's N), where a letter string's N size is the number of real words that can be created by substituting a single letter, such as *hat*, *cut* and *cap*, which are all neighbours of the word *cat* (M. Coltheart et al., 1977). A more recent approach is based on the Orthographic Levenshtein Distance (OLD): here, for each letter string, the most similar words are derived, and the number of letter deletions, additions, or substitutions of the twenty closest words are averaged to get an OLD20 measure (Yarkoni et al., 2008).

Our paper specifically concerns a measure called body-N, which is a theory-driven alternative method of capturing similarity to real words (Ziegler & Perry, 1998). For monosyllabic words, a particularly psychologically salient orthographic unit is the body, which consists of the syllable's vowel and coda, such as the cluster *-ird* in the word *bird* (Forster & Taft, 1994; Goswami, 1999; Goswami & Bryant, 1990; Treiman, Goswami, & Bruck, 1990; Treiman et al., 1995; Ziegler & Goswami, 2005). Body-N is calculated, for a given letter string, as the number of words that have the same body, regardless of the size or overlap of the onset (initial consonant cluster). For example, the nonword *lat* has 16 body neighbours, including the words *at*, *cat*, *brat* and *sprat.*

The broad question addressed in the current data review is the extent to which word and nonword reading are influenced by the number of body neighbours. Previous studies have generally compared reaction times of words and nonwords with high body-N to items with low body-N (Ziegler & Perry, 1998; Ziegler et al., 2001; Ziegler, Perry, Ma-

Wyatt, et al., 2003). Note that here we are concerned with studies which compare a high body-N condition to a low body-N condition. Studies comparing the reading of nonwords with existing versus non-existing bodies in reading aloud (e.g., *dake* - *daik*) are more numerous and consistently show facilitation associated with body existence (e.g., Andrews, Woollams, & Bond, 2005; Goswami, Ziegler, Dalton, & Schneider, 2003; Rosson, 1985; Treiman et al., 1990), but this design cannot be used to address theoretically important questions about interactions between body-N and lexicality or word frequency, because a real word cannot have a non-existent body.

Although a great amount of research has been dedicated to understanding the effects of orthographic or Coltheart's N (e.g., see Andrews, 1997, for a review), it is still unclear what mechanisms drive body-N effects in reading. Nevertheless, research suggests that bodies are psychologically salient units (Forster & Taft, 1994; Goswami & Bryant, 1990; Schmalz et al., 2014; Treiman et al., 1995; Ziegler & Perry, 1998; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). Therefore, it is important to further explore the stability of the body-N effect, as well as theoretically interesting interactions, to understand the origins of this effect within the workings of the reading system.

2.2. Theoretical relevance

2.2.1. Are bodies processed by a lexical or sublexical route?

Although several studies have been conducted on body-N effects (Ziegler & Perry, 1998; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003), it is still unclear what finding a body-N effect actually means. Assuming a dual-route framework, there are two broad possibilities: either bodies reflect processes that are occurring in the lexical route, or they reflect functioning of the sublexical route (or both). If it is possible to establish whether bodies are processed by the lexical or sublexical route, this would provide a new marker effect that would need to be simulated by computational models.

This is important, because the classical computational implementation of the dual-route theory, the dual-route cascaded (DRC) model (M. Coltheart et al., 2001) has been criticised for its insensitivity to units that are larger than graphemes (M. Coltheart, 2012; Jared, 2002; Perry et al., 2007; Treiman et al., 2003).

The first possibility is that bodies reflect lexical processing. Here, body-N effects would be driven by activation between lexical nodes, which could be easily accounted for by an interactive activation model (e.g., Forster & Taft, 1994; Taft, 1991). In this view, finding a body-N effect would show that the orthographic space is organised in a hierarchical manner, such that words that resemble each other by having the same body are co-activated with the target to a greater extent than words that have the same degree of orthographic overlap but involve different units (Forster & Taft, 1994).

If we assume that bodies are processed by a sublexical route, there could be a simple way of increasing the DRC's sensitivity to larger orthographic units, namely to insert body-rime correspondences into its sublexical route (M. Coltheart, 2012; Patterson & Morton, 1985). A more recent implementation of the dual-route model, the CDP+ (Perry et al., 2007), provides one possibility of creating a sublexical route that shows sensitivity to marker-effects associated with large units, such as body-N and body consistency effects. Body consistency relates to the number of possible pronunciations of a body: the body *-eek* is consistent, as it only has one possible pronunciation, while the body *-eak* can be pronounced as in "leak" or "break" and is therefore inconsistent (see also previous paper). Furthermore, the body consistency effect has been shown by the authors to be driven by the sublexical route: eliminating all feedback activation of the lexical route (which drives lexical similarity effects) does not change the CDP+'s results pattern when it comes to simulating the consistency effect.

The sublexical route of the CDP+ differs from that of the DRC in that the sublexical correspondences are learnt via a two-layer associative network. This allows the sublexical network to develop sensitivity to context-specific regularities of the orthography-phonology conversion, which includes taking into account the coda to predict vowel pronunciation (cf. Treiman et al., 1995).

Similarly, triangle (PDP) models contain an orthography-phonology conversion route which is based on a connectionist network (Harm & Seidenberg, 2004; Plaut, 1999; Plaut et al., 1996). Here, the system learns the correspondences that underlie the orthography's print-to-speech regularities by taking into account larger grain-sizes, when smaller grain-sizes (such as letters or graphemes) are unreliable. Therefore, such models might become sensitive to body-rime correspondences due to the pressure to minimise pronunciation ambiguity in the orthography-phonology conversion procedure.

In summary, if we establish that bodies are processed in the sublexical route, this could either be simulated by a connectionist sublexical route, which learns the regularities between print and speech while taking into account the surrounding letters, or by a rule-based sublexical route, if we insert body-rime correspondences in addition to grapheme-phoneme correspondences.

Given the current state of knowledge, however, is not yet clear whether bodies are processed by a lexical or a sublexical route. This is an empirical question, which can, in theory, be addressed by a thorough examination of body-N effects and its interactions with task, lexicality, and frequency. Marker effects associated with sublexical processing, such as the length or word regularity effect, have been shown to interact with lexicality and frequency, such that they are stronger for nonwords than words, and for low-frequency words than high-frequency words (Cummine, Amyotte, Pancheshen, & Chouinard, 2011; Hino & Lupker, 2000; Paap & Noel, 1991; Weekes, 1997; Ziegler et

al., 2001). In a dual-route framework, this occurs because the lexical route operates quickly for a high-frequency word, thus suppressing the impact of the sublexical route in responding to this particular word.

Therefore, if body-N reflects a sublexical mechanism, we should expect, in reading aloud latencies, a stronger body-N effect for nonwords than for words, and an interaction between body-N and frequency for words, where the strength of the effect decreases with increasing frequency. If body-N reflects a lexical mechanism, conversely, we might expect stronger body-N effects for words than for nonwords.

If the body-N effect reflects a sublexical mechanism, we can also make predictions about how it would behave in a lexical decision task: lexical decision is hypothesised to require less sublexical processing than reading aloud, as the reliance on lexical activation does not technically require the input of a sublexical route (Hino & Lupker, 2000; Schmalz, Marinus, & Castles, 2013). Therefore, we may expect a reduced body-N effect in lexical decision compared to the reading aloud tasks, or no body-N effect at all. If the body-N effect reflects a lexical mechanism, we might expect, for words, a stronger body-N effects for lexical decision than for reading aloud. For nonwords in lexical decision, we might expect an inhibitory effect, if the overlap between the target and its existing neighbourhood words activates lexical nodes that make it harder for the nonword to be rejected (as has been previously found for orthographic N; see M. Coltheart et al., 1977).

2.2.2. Body-N effects across languages

The body-N effect has been used as a marker effect of reliance on body units across languages (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). It has previously been proposed that the degree to which readers rely on body units depends on the depth of their orthography (Goswami, 1999; Ziegler & Goswami, 2005). "Deep"

orthographies are defined as those having unreliable sublexical correspondences. For example, in English – a notoriously deep orthography (Share, 2008) – the letter string *ough* has different pronunciations in the words *through, tough, though, bough, cough,* and *hiccough* (Ziegler et al., 1997). Corpus analyses have shown that some of the inconsistencies in English can be resolved via the use of body-units (Peereman & Content, 1998; Treiman et al., 1995): for example, although the pronunciation of the word *palm* does not correspond to its individual letter-sound correspondences ("/pælm/"), its body pronunciation can be predicted by analogy to words like *calm* or *balm*.

According to the psycholinguistic grain size theory (Ziegler & Goswami, 2005), a prominent theory of reading development across languages, the lack of reliability of sublexical correspondences in deep orthographies, – along with the finding that larger units tend to be more predictive of a word's pronunciation than small units, – forces readers to develop sensitivity to larger units. Conversely, readers of "shallow" orthographies can achieve accuracy while relying on small units (i.e., letters, graphemes) only. This is hypothesised to result in different decoding strategies during childhood, which leave footprints in the cognitive processes underlying reading in adults (Ziegler et al., 2001).

Therefore, a body-N effect should be stronger in deep compared to shallow orthographies – and indeed, there is evidence supporting this in two cross-linguistic studies comparing single word and nonword reading aloud in English, a deep orthography, and German, a shallow orthography, with adults (Ziegler et al., 2001) and children (Ziegler, Perry, Ma-Wyatt, et al., 2003). However, directly comparing the size of the body-N effect across languages in an orthogonal design requires a great deal of methodological rigour in order to avoid alternative explanations due to confounding variables that differ systematically across the two languages. The English orthography,

for example, is denser than the German orthography. This means that words, on average, have more neighbours, including body-neighbours, in English than German. In the case of the item set of Ziegler et al. (2001; 2003), this leads to a systematically higher body-N count for English high body-N items (mean = 15.33, SD = 6.00) compared to the German high body-N items (mean = 10.28, SD = 4.35). The low body-N items did not differ across languages in terms of the average body-N size (mean = 4.66, SD = 3.75; mean = 4.29, SD = 3.75, in English and German respectively). This means that the size of the body-N manipulation was stronger for English than German, with the differences of body-N size between the high and low body-N conditions of 10.67 and 7.99 respectively. This is problematic when drawing conclusions about the degree of reliance on body units across languages as a function of orthographic depth, as the stronger manipulation offers a viable alternative explanation.

In summary, open questions remain about the size of the body-N effect across orthographies varying in depth, because the interpretation of currently available data is limited by uncontrolled cross-linguistic differences. As the data on the body-N effect across languages is relatively sparse, it is desirable to replicate the findings of Ziegler et al. (2001). A large-scale meta-analysis is particularly useful for this end. Firstly, it will allow us to treat body-N as a continuum, and thereby eliminate the bias associated with the different size of the manipulation of the conditions. Secondly, it will allow us to assess the stability of the body-N by language interaction, while taking into account potential confounds, such as orthographic N, as a covariate.

2.3. Methodological considerations

We aim to address the theoretical questions underlying body-N effects and interactions by assessing these in a large scale meta-analysis of all available studies of reading which used body-N manipulations. In addition to allowing us to have a large item

set and large sample size, this approach has several advantages. Firstly, body-N is a naturally continuous variable. When controlling for potential confounds in an orthogonal design, the pool of items becomes very small. As a result, body-N has been dichotomised in previous studies, because the small number of items generally used gives insufficient power to treat body-N as a continuous predictor. This means that a reliable effect size of the body-N effect cannot be established, as it may vary across studies along with the strength of the manipulation. With a meta-analysis including data from several studies, we can address this issue by using body-N as a continuous predictor of RTs.

Secondly, increasing the power by using a meta-analysis will also give us more reliable estimates of theoretically meaningful interactions. As described above, we are interested in the following: (1) The stability of main effects of body-N, (2) interactions between frequency and body-N, (3) differential body-N effects for words and nonwords, (4) differential body-N effects for reading aloud and lexical decision, and (5) interactions between body-N and language.

Thirdly, designing a well-controlled psycholinguistic experiment is near impossible, given the inter-correlated nature of most linguistic properties within a lexical corpus (e.g., Andrews, 1997; Cutler, 1981; Kliegl, Grabner, Rolfs, & Engbert, 2004). For example, body-N co-varies systematically with orthographic N: a large number of body-neighbours of a given word are also its orthographic neighbours. Therefore, a random item with low body-N is also likely to have few orthographic neighbours, and a random item with high body-N is likely to have many orthographic neighbours. During item selection, it is possible to counteract this by picking high body-N items with complex onsets, as these tend to have fewer orthographic neighbours. This, however, inflates the number of letters and the syllabic complexity (due to complex onset clusters) of the high body-N condition. This, in turn, may diminish or mask theoretically important effects and

interactions. We circumvent this problem by taking into account covariates, such as orthographic N, number of letters, and the number of consonants in the onset – which is possible given a large enough item set and sample size.

Furthermore, as discussed in the previous section, the issue of confounds complicates the conclusions that may be drawn from cross-linguistic studies. In relation to previous work on body-N as a function of orthographic depth, there are systematic differences in the orthographic density between English and German; consequently, differences in the size of the body-N effect found by Ziegler et al. (2001; 2003) may be either attributed to orthographic depth, or to a stronger body-N manipulation for English than German. Using body-N as a continuum allows us to circumvent this problem, as each item is considered in relation to its individual body-N value, rather than the average of an experimentally designed condition.

2.4 Method

2.4.1. Studies included in the analyses

We analysed all available skilled adult readers studies which used either single word reading aloud or lexical decision and manipulated the number of body-neighbours for words and/or nonwords. Altogether, we know of 11 studies that have included such manipulations. Two studies have been published by Ziegler and colleagues: one used a lexical decision task in English (Ziegler & Perry, 1998), and the other used a reading aloud task with English and German readers (Ziegler et al., 2001). For these two studies, the by-trial data (i.e., RT data which have not been averaged across items or participants) have been lost (J. Ziegler, personal communication, 2013 & 2014). For the Ziegler et al., (2001) study, the item-level data (i.e., the average for each item collapsed across participants) is available in the appendix of another paper (Perry & Ziegler, 2002).

Another study on the body-N effect has been conducted in English by Taft (unpublished; personal communication, 2014). This study used a lexical decision task with two sets of items: for the first set, words were manipulated on body-N while orthographic N was held constant, while in the second set, orthographic N was manipulated, while body-N was held constant. This study has not been published, because the results showed only marginally significant effects, which were not stable in either across-items or across-subjects analyses. For this study, the trial-level data are available. Eight further studies have been conducted as part of this dissertation. These included both lexical decision and reading aloud tasks, in both English and German (see Appendix).

The overall item characteristics across all studies that were included in the analyses (averages, SDs and correlations with body-N) are described in Table 1. The body-N counts are based on the same corpus analysis as those of Ziegler et al., (2001) to increase the comparability across languages (Ziegler et al., 1997, for English, and Ziegler, 2012, personal communication, for German). The frequency and orthographic N values are taken from WordGen (Duyck et al., 2004), which is an interface for cross-linguistic research based on the CELEX database (Baayen et al., 1995). Regularity was defined as compliance to the GPC rules, as implemented in the German and English versions of the DRC (M. Coltheart et al., 2001; Ziegler et al., 2000).

2.4.2. Separate analyses of previous studies

We outline the characteristics and basic results of the individual studies, as per a re-analysis using Linear Mixed Effect (LME) models (Baayen, 2008; Baayen et al., 2008) in the Appendix. Note that the table includes only the studies for which we had available trial-level data (meaning that, for the time being, we exclude the two published studies by Ziegler and colleagues), as we need trial-level data to include subjects and items as

random effects. The results of the two studies by Ziegler and colleagues are discussed in detail in a later section.

The *t* and *p* values analyses in the "Results" column are based on LMEs, using body-N as a continuous predictor for inverse RTs (-1000/RT), with items and subjects as random effects. The Bayes Factor (BF) analyses were conducted with the BayesFactor package for R (Morey & Rouder, 2014). In the analyses of Table 2, BFs exceeding 3 are considered to provide evidence for the presence of a body-N effect (H_1); BF values between 1/3 and 3 provide equivocal evidence for the H_1 and H_0, and BF values smaller than 1/3 provide evidence for the absence of the body-N effect (H_0) over H_1 (Rouder et al., 2009).

The absence of an effect as shown by a non-significant *p*-value is hard to interpret, as it cannot distinguish between the possibility that there is no effect and the possibility that the data is insensitive to picking up an effect. This problem can be circumvented when we use BFs, as these can distinguish between these two scenarios. Specifically, the BF value can tell us when the data is more strongly in support of H_0 than H_1, or whether the data shows equivocal evidence for H_0 and H_1. We therefore inspected the BF values as a function of task, lexicality, and language. These are summarised in Table 2.

Table 2 shows that for most types of items, the BF provides either equivocal evidence for H_1 and H_0, or stronger evidence for H_0 than H_1. This suggests that it is unlikely that there are body-N effects for those types of items. Two exceptions to this are the results for the lexical decision task for German nonwords, which provide solid evidence for H_1 (an inhibitory effect of body-N), and reading aloud for English nonwords, where each of the three studies provides a different outcome ($H_1 > H_0$, $H_1 \approx H_0$, and $H_1 < H_0$).

Table 1.

Descriptive statistics across all items included in the analyses, and their correlation with body-N. Corr = correlation between body-N and the given variable.

	English						German					
	Words (N = 283)			Nonwords (N = 305)			Words (N = 158)			Nonwords (N = 169)		
	Range	Mean (SD)	Corr.	Range	Mean (SD)	Corr.	Range	Mean (SD)	Corr.	Range	Mean (SD)	Corr.
Body-N	1-23	8.65 (5.96)	N/A	0-20	7.91 (5.48)	N/A	1-19	6.92 (5.41)	N/A	0-19	6.58 (5.21)	N/A
Number of letters	3-7	4.47 (0.88)	r = 0.01, p = 0.87	3-6	4.61 (0.78)	r = -0.02, p = 0.78	3-6	4.38 (0.93)	r = 0.10, p = 0.22	3-7	4.40 (0.91)	r = 0.16, p < 0.05
Orthograph. N	0-25	5.43 (4.54)	r = 0.29, p < 0.001	0-24	4.41 (4.17)	r = 0.17, p < 0.01	0-15	3.72 (3.02)	r = 0.16, p < 0.05	0-15	3.59 (2.79)	r = 0.14, p = 0.07
Onset complexity	0-3	1.57 (0.61)	r = 0.34, p < 0.001	1-3	1.71 (0.53)	r = 0.31, p < 0.001	0-4	1.39 (0.75)	r = 0.20, p < 0.05	1-4	1.47 (0.66)	r = 0.40, p < 0.01
Consistency ratio	0.25-1	0.95 (0.12)	r = 0.01, p = 0.92	0.5-1	0.97 (0.08)	r = 0.02, p = 0.76	0.08-1	0.97 (0.12)	r = -0.02, p = 0.79	0.5-1	0.98 (0.08)	r = 0.02, p = 0.81
% Regular words	N/A	94.40%	N/A	N/A	N/A	N/A	N/A	96.05%	N/A	N/A	N/A	N/A
Log frequency	0.00-3.92	1.59 (0.67)	r = -0.02, p = 0.68	N/A	N/A	N/A	0.00-4.40	1.49 (0.72)	r = 0.12, p = 0.13	N/A	N/A	N/A

In summary, taken together, the results from all available unpublished studies for which there is available trial-level data are not consistent with the view that body-N has an influence on reading, except for the lexical decision task for nonwords in German, where there is stable evidence for an inhibitory body-N effect (albeit based on only one study), and the reading aloud task for nonwords in English, where the evidence is mixed.

Table 2.

Summary of previous results: Numbers of studies in each category. BF > 3 = number of studies providing support for a body-N main effect, BF ≈ 3 is equivocal evidence for and against the main effect of body-N, and BF < 1/3 is the number of studies providing evidence against an influence of body-N.

	English				German			
	Reading aloud		Lexical decision		Reading aloud		Lexical decision	
	Words	Non-words	Words	Non-words	Words	Non-words	Words	Non-words
BF > 3	0	1	0	0	0	0	0	1
BF ≈ 1	0	1	1	1	1	0	0	0
BF < 1/3	2	1	2	2	1	2	1	0

2.5. Analyses and Results

In the subsequent section, we further analyse all available data. We first conduct a meta-analysis which includes all studies for which there is trial-level data. As this happens to exclude the published studies of Ziegler and colleagues, we describe these in a separate section.

2.5.1. Meta-analyses

To further explore the pattern of results, we collapsed across all studies with available trial-level data to assess the stability of the effects to obtain greater power. As some of the BF values from Table 2 and the Appendix showed equivocal evidence for an

influence of body-N, increasing the item and sample size by collapsing across a number of studies can be used to draw more confident conclusions about the stability of body-N effects.

We performed four groups of analyses: for nonwords in reading aloud, nonwords in lexical decision, words in reading aloud and words in lexical decision. We analysed these conditions separately, because different cognitive mechanisms underlie response latency variance in each of the four conditions. This should be reflected in different patterns of the body-N effect. For example, we expect a facilitatory body-N effect for reading aloud of words and nonwords, and for lexical decision for words, as stronger activation of a body unit may enhance lexical activation and/or the sublexical assembly process (Ziegler & Perry, 1998; Ziegler et al., 2001). For lexical decision for nonwords, however, we might expect an *inhibitory* body-N effect: if a high body-N nonword elicits more lexical activation compared to a low body-N nonword, it will be harder to reject in the lexical decision task. Each analysis included both the English and German items, which enabled us to assess any interactions between body-N and language, as this is relevant for the Psycholinguistic Grain Size Theory (Ziegler & Goswami, 2005).

We analysed each type of items with LMEs and BFs. LME models are used widely in psycholinguistic research, as they can handle the interdependence of subject- and item-level variance by including these as random factors. We sought to find converging evidence from the BF approach: Firstly, to ensure that the findings are reliable, and secondly, to be able to interpret null-results.

2.5.2. Measuring body-N: Types versus tokens

The two published studies of the body-N effect used type body-N, or the number of words with the same body to quantify the effect (Ziegler & Perry, 1998; Ziegler et al., 2001). In the literature on word consistency effects, some evidence suggests that reliance

on large units is instead driven by token frequency (Jared et al., 1990), which can be quantified, in the context of the body-N effect, as the summed frequency of all body neighbours. Practically, type and token counts are difficult to dissociate unless the item sets are created with the aim of de-correlating these variables, due to a high correlation between them (e.g., $r(915) = 0.43$, $p < 0.0001$, in the items included in this analysis).

At the beginning of each set of analyses, we compared models including type versus token body-N as predictors. Our aim here was to isolate the more reliable predictor rather than adjudicating between the two measures. Each model comparison showed a numerical advantage for type body-N compared to token body-N according to measures of model fit, as indicated by the Akaike Information Criterion (AIC). For this reason, and also because previous research has used type body-N counts to quantify the body-N effect, we use type body-N counts for all subsequent analyses.

2.5.3. Body-N effect for nonwords in reading aloud

For two reasons we might expect the most interesting results for the reading aloud nonwords condition. Firstly, if reliance on bodies reflects a sublexical strategy (M. Coltheart et al., 2001; Patterson & Morton, 1985; Perry et al., 2007), we would expect to find the strongest body-N effect for this condition. If this is the case, this condition will also give us the most leverage for assessing the interaction between body-N and language. Secondly, the three English studies on reading aloud nonwords, as summarised in Table 3, give conflicting results about the existence of the body-N effect: for the first study, the BF provides evidence for an effect, in the second there is equivocal evidence for the body-N effect and the H_0, and the third provides evidence against the presence of a body-N effect.

The analyses were conducted on inverse RTs as the dependent variable. The predictors were body-N, language (German, English - dummy coded as 1 and -1,

respectively, in the LME analyses, in order to obtain estimates of the main effects of language as a deviation from the grand mean), orthographic N, and onset complexity (the number of consonants in the onset)[13]. The continuous predictors were centred by subtracting their mean from each value, so as to obtain LME parameter estimates for average values rather than extreme values of zero. We also included subject, item, and study as random factors in all analyses.

2.5.3.1. LME model analysis

Comparing models containing no interactions, all two-way interactions, and the three-way interaction between body-N, orthographic N, and language, we found that the model containing two-way interactions performed significantly better than the model with no interactions, χ^2 (3) = 18.41, $p < 0.001$, while there was no additional benefit of adding three-way interaction, χ^2 (1) < 1.

In the model including the main effects and two-way interactions between body-N, orthographic N and language, as well as the main effect of onset complexity, we found a facilitatory main effect of orthographic N, an interaction between body-N and language (which we elaborate on in detail below), and an interaction between language and orthographic N, due to a stronger facilitatory orthographic N effect in German than English. The interaction between body-N and orthographic N approached significance (as we discuss below). The main effect of body-N did not approach significance. The LME results are summarised in Table 3.

[13] Onset complexity is mainly included to act as a covariate. Mostly, the studies that were included in the analyses manipulated body-N while keeping orthographic N constant, meaning that high body-N words tended to contain more complex onset clusters to reduce the orthographic N value. As this may act to suppress a body-N effect, we included the effect of onset complexity as a statistical control.

Table 3.

Output from the LME analysis for reading aloud nonwords, including body-N, Language, Orthographic N, and their two-way interactions, and the main effect of onset complexity.

	Estimate	Std. Error	t value	p value
Intercept	*-1.741*	*0.036*	*-48.421*	*<0.00001*
Main effects				
Body-N	-0.001	0.001	-0.401	0.689
Language	-0.008	0.029	-0.293	0.770
Orthographic N	-0.013	0.002	-6.425	<0.00001
Onset complexity	-0.019	0.014	-1.412	0.159
2-way interactions				
Body-N x language	0.003	0.001	3.194	0.001
Body-N x orthographic N	<0.001	<0.001	1.702	0.089
Language x orthographic N	-0.004	0.002	-2.593	0.010

We further explored the patterns of interactions in the results. Specifically, we sought to examine the interaction between language and body-N, and the marginally significant interaction between body-N and orthographic N. The former is theoretically important, as establishing whether we find a reliable interaction between language and body-N was one of our aims. Given the absence of the main effect of body-N, it is important to establish whether body-N effects may be modulated by other variables. If there a body-N effect were to emerge only for nonwords with particular characteristics, this might explain why there is conflicting evidence for and against the presence of an effect according to the BFs, as shown in Table 2.

To assess the source of possible interactions between body-N and language and orthographic N, we obtained the body-N slope estimates (1) for English high orthographic N items, (2) for English low orthographic N items, (3) for German high orthographic N

items, and (4) for German low orthographic N items. High orthographic N was defined as 1 SD above the mean, and low orthographic N as 1 SD below the mean.

These analyses showed the following pattern: (1) For English, high orthographic N, there was a marginally significant facilitatory body-N effect, slope = -0.003, t = -1.75, p = 0.08. (2) For English, low orthographic N, there was a strong facilitatory body-N effect, slope = -0.005, t = -2.74, $p < 0.01$. (3) For German, high orthographic N, there was a marginally significant inhibitory trend of a body-N effect, slope = 0.004, t = 1.95, p = 0.05, and (4) for German, low orthographic N, there was a small, but not significant trend for an inhibitory body-N effect, slope = 0.002, t = 0.075, p = 0.45.

This suggests that if body-N has a real effect on nonword reading aloud latencies, it does so via a complex interactive pattern, where the effect that it has is dependent on both the language and the orthographic N of the item. In particular, the body-N by language interaction, in the absence of an overall main effect of body-N, appears to be driven by a small inhibitory body-N effect in German, and a small facilitatory effect in English. High orthographic N appears to be associated with stronger facilitation of high body-N in both languages.

2.5.3.2. Bayes Factor analysis

Due to the post-hoc and exploratory nature of the current analyses, and the non-significant p-value associated with the main effect of body-N, it is difficult to draw any strong conclusions. We therefore sought converging evidence from an alternative approach, namely a BF analysis. As described in the introduction, a BF analysis can provide either evidence for a model compared to the model that it is tested against, if the BF value is larger than 3. BF values smaller than 1/3 provide evidence against the model that is being tested, and values between 1/3 and 3 are considered to provide equivocal evidence for the model (Rouder et al., 2009).

To mirror the LME analyses, we started with a comparison of a main-effects model (including language, orthographic N and body-N, as well as onset complexity as a covariate and items, participants, and study as random effects) to one which also included all two-way interactions. This provided evidence against the main-effects only model, BF = 0.11 (±1.93%). We further compared this two-way interactions model to a model including the three-way interaction, and - again - found evidence for the two-way interaction model, BF = 5.43 (±1.73%). We therefore adopted the two-way interaction model as a baseline for further model comparisons.

To establish the importance of the main effect of body-N, we compared the two-way interactions model to one excluding both the main effect of body-N, and any interactions associated with it. Here, BF = 0.42 (±3.55%), thus providing weak evidence against any influence of body-N - though it does not go below the conventionally critical cut-off point of 1/3, which would indicate evidence against the model including body-N.

Even though the BF analysis does not favour a model which includes both the effects and interactions of body-N, it is not clear that we can conclude that there is neither a main effect, nor interactions of body-N. It is possible, for example, that the main effect of body-N improves the model fit, but including the interactions decreases it and thereby counteracts a meaningful main effect. We therefore followed up with further model comparisons to establish the importance of the relevant effect and interactions.

To assess the importance of the main effect of body-N, we compared the "base" model (Language and orthographic N and their interaction, plus main effect of onset) to one which also included the main effect of body-N. For the model including the main effect of body-N, BF = 0.19 (±3.55%), suggesting that as a main effect, body-N is unlikely to have any influence on reading aloud nonword latencies.

As this does not rule out the possibility of a body-N by language interaction, which was significant in the LME analysis, we compared the model which included body-N (same as the body-N model for the previous analysis) against one which also included the interaction between body-N and language. Here, we found support for the model which included the interaction: BF = 5.07 (±1.87%).

The final model served to follow up on the marginally significant interaction between body-N and orthographic N in the LME analysis. If we find further evidence for the presence of such an interaction, future research might need to follow up with an independent investigation of this finding. If we do not find evidence for this interaction, this will indicate that it is likely to reflect a spurious finding. Against the full, two-way interaction model, we compared one which was identical except that it excluded the interaction between body-N and orthographic N. Here, we found equivocal evidence for the model which included the interaction of body-N and orthographic N, BF = 2.25 (±1.75%). Given this inconclusive result, and the fact that this interaction was only marginally significant in the LME analyses, we take a conservative approach and assume that the interaction is likely to be a spurious finding, and do not discuss it further.

2.5.3.3. Summary

The original aims were to establish whether there is a main effect of body-N, and whether body-N interacts with language. Both in the LME and BF analyses, we found no evidence for the presence of a main effect of body-N. The interaction between language and body-N emerges consistently in all analyses; however, it appears to be driven by a pattern that is unlikely to hold true. As shown in the analyses of body-N slopes for English compared to German, English readers showed a trend towards a facilitatory effect of Body N, while Germans showed a trend towards an inhibitory effect. In the LME analyses, the body-N slope was significant only for low orthographic N English

nonwords, and showed non-significant trends at best at all other points that we tested. This apparent cross-over seems to be driving the significant body-N by language interaction. We are inclined to dismiss the result as a Type I error, since we know of no model, theory, or other dataset that would suggest an inhibitory body-N effect for German, but facilitation for English. However, should this result prove replicable in other studies, this position and extant theories would have to be revisited.

It is noteworthy that the LME analysis showed an interaction between orthographic N and language (see Table 3). This is due to a stronger orthographic N effect in German (slope = -0.02, t = -5.40) than in English (slope = -0.01, t = -4.20). To follow up, we conducted a BF analysis comparing the full two-way interaction model described above to one excluding the language by orthographic N interaction. In line with the LME results, this showed evidence for the presence of this interaction, BF = 3.08 (±2.83%). For German and for English separately, both BFs provide very strong evidence for the presence of an orthographic N effect, for German BF = 17,045 (±1.50%), and for English BF = 187 (±2.22%). Together, these findings suggest that orthographic N has a facilitatory effect on both the German and English reading aloud latencies, but this effect is stronger for German than English. Although we have no straightforward explanation for this result (and it is not related to our original aims), it has been previously suggested that cross-linguistic differences might account for some contradictory results in the literature on orthographic N (Andrews, 1997). Therefore, this result may be of interest to researchers seeking to understand cross-linguistic differences in the size of the orthographic N effect.

2.5.4. Body-N effects for nonwords in lexical decision

We performed an equivalent set of analyses for the nonwords in the lexical decision task. The dependent variable and independent variables were identical to the previous set of analyses.

2.5.4.1. LME model analysis

We found no advantage of any model including interactions over one containing main effects only based on measures of model fit, both $\chi^2 < 4$ and $p > 0.2$. The main-effects only model for type body-N showed an inhibitory effect of body-N, $t = 2.86$, $p < 0.005$, and an inhibitory effect of orthographic N, $t = 3.23$, $p < 0.005$. All other $p > 0.4$.

2.5.4.2. Bayes Factor analysis

We compared the main effects only model to one which included two-way interactions, and to one which included three-way interactions. In both cases, the evidence was in favour for the main-effects only model, BF > 100, which is adopted for further comparisons.

A BF comparison of the full main-effects model compared to one which excluded the main effect of body-N provided support for the H_1, that body-N contributes to lexical decision latencies: BF = 4.75 (±2.66%). A comparison of the model which included an interaction between language and body-N as well as the main effects (H_1) against a main-effects-only model (H_0) provided evidence against the interaction, BF = 0.19 (±2.50%).

2.5.4.3. Summary

A relatively simple model that included no interactions was supported in the current set of analyses. We found a stable inhibitory body-N effect for nonwords in lexical decision, in addition to an inhibitory effect of orthographic N. There was evidence against an interaction with language.

2.5.5. Body-N effects for words in reading aloud

In the third set of trial-level analyses, we explored the effects and interactions of body-N in the reading aloud task for words. The dependent and independent variables were identical to those for nonwords, but frequency was included as an additional predictor. An interaction of body-N and frequency is theoretically important: If bodies are processed as sublexical units, we should find a smaller effect for high-frequency words, because the rapid lexical activation associated with the processing of high-frequency words would mask the sublexical effect.

2.5.5.1. LME model analysis

Initially, we compared models with no interactions, to models also including two-way or three-way interactions or the four-way interaction. A model including the four-way interaction (between body-N, frequency, orthographic N and language) was favoured over the three-way interaction model, χ^2 (1) = 4.22, $p < 0.05$. The results of the full LME model including four-way interactions are summarised in Table 4.

Table 4.

Output from the LME analysis for reading aloud words, including body-N, Language, Orthographic N, frequency, and the interactions, and the main effect of onset complexity.

	Estimate	Std. Error	t value	p value
Intercept	*-1.904*	*0.032*	*-59.046*	*<0.001*
Main effects				
Body-N	0.002	0.001	1.266	0.206
Language (German)	-0.011	0.054	-0.205	0.838
Orthographic N	-0.003	0.002	-1.336	0.182
Log frequency	-0.034	0.012	-2.884	0.004
Onset complexity	-0.061	0.012	-5.169	<0.001
2-way interactions				
Body-N x language	< 0.001	0.002	0.109	0.913
Body-N x orthographic N	-0.001	< 0.001	-2.213	0.028
Body-N x log frequency	-0.003	0.002	-1.693	0.091
Language x orthographic N	0.001	0.004	0.316	0.752
Language x log frequency	0.043	0.021	2.002	0.046
Orthographic N x log frequency	< 0.001	0.004	0.039	0.969
Three-way interactions				
Body-N x language x orthographic N	-0.001	0.001	-1.641	0.102
Body-N x language x log frequency	0.008	0.004	1.910	0.057
Body-N x orthographic N x log frequency	< 0.001	0.001	0.217	0.829
Language x orthographic N x log frequency	0.003	0.007	0.367	0.714
Four-way interaction				
Body-N x language x orthographic N x log frequency	-0.002	0.001	-2.027	0.043

The analysis showed a facilitatory main effect of frequency, and a facilitatory main effect of onset complexity. The two-way interaction between language and frequency occurred due to a stronger frequency effect for English than for German. There was a significant interaction between body-N and orthographic N. Analogous to the reading aloud nonword analyses, we followed up with an estimation of the body-N slope at four different points: (1) for German low orthographic N words (i.e., for orthographic N values which are 1 SD below the mean); (2) for German high orthographic N words (1 SD above the mean); (3) for English low orthographic N words, and (4) for English high orthographic N words.

For German, the body-N slope was significant at low orthographic N values, $t = 2.61$, $p < 0.01$, indicating an inhibitory body-N effect. At high orthographic N values, the body-N slope showed was inhibitory, but not close to significant, $t < 1$. For English, the body-N slope was not significant either for low or high orthographic N words, for low orthographic N, $t = 1.07$, $p > 0.2$, and for high orthographic N, $t < 1$. In both cases, the directions of the body-N slope indicated an inhibitory trend. Thus, the interaction between body-N and orthographic N seems to reflect an overall stronger inhibitory body-N effect for high orthographic N items compared to low orthographic N items, though the only point at which the body-N effect became significant was at German low orthographic N.

The two-way interaction between body-N and frequency was marginally significant, therefore we similarly followed up with estimations of the body-N slopes for different points across language and frequency. For German, the body-N effect was not significant at either low-frequency words (1 SD below the mean frequency), nor at high-frequency words, both $t < 1$.

For English, at low-frequency words the body-N slope reached significance, and showed an inhibitory body-N effect, $t = 2.04$, $p < 0.05$. For high-frequency words, the body-N slope was not significant, but showed a facilitatory trend, $t = -1.16$, $p > 0.2$.

The four-way interaction was significant. This is likely to be driven by the pattern described above: the body-N effect only emerges as being significant for low orthographic N German words, and shows a cross-over from facilitatory to inhibitory across frequency for English words. The slopes tend to indicate an inhibitory body-N effect, which is in contrast to previous findings, and to the predictions that we outlined in the introduction.

2.5.5.2. Bayes Factor analysis

In contrast to the LME analyses, the BF analysis did not show support for any of the interaction models over a main-effects only model, all BF > 9000. Therefore, the model used in the following BF analyses included only the main effects of body-N, orthographic N, frequency, and language, as well as onset complexity as a covariate and study, item, and subject as random factors.

To establish whether body-N had an effect of reading aloud latencies, we compared a main effects model which excluded the body-N effect to one which included it. Here, we obtained evidence against the presence of a main effect of body-N, although this missed the 1/3 benchmark, BF = 0.42 (±1.78%).

There are two interactions which are theoretically important according to the predictions that we outlined in the introduction: firstly, the body-N by language interaction, and secondly the body-N by frequency interaction. We therefore compared main effect models which also included each of the interactions, to the main-effects only model. We obtained evidence against the model which includes the body-N by language

interaction, BF = 0.23 (±2.76%), and against the model which includes the body-N by frequency interaction, BF = 0.14 (±1.75%).

2.5.5.3. Summary

The LME and BF approach give diverging results about the interactions underlying the body-N effect for reading aloud words. The LME suggests that both frequency and orthographic N differentially mediate the size of the body-N effect, such that it becomes significantly inhibitory in German for low orthographic N words, and in English for low frequency words. The BF, however, very strongly supports a main-effects only model, without any interactions, suggesting that the LME pattern is driven by spurious results. In addition to the evidence from the BF analysis, we are inclined to place little weighting on the interactive pattern of the LMEs because there is no *a priori* reason for expecting an inhibitory body-N effect, nor the specific interactive patterns that we found.

Overall, the results of both the LMEs and the BF are in line with a view that there is no body-N effect, and no body-N by language interaction. Concerning the theoretically important interaction between body-N and frequency, the LME and BF approaches disagreed: the LME showed an interaction with frequency while the BF provided evidence against it. For the reasons described above, we adopt the view that frequency does not mediate the body-N effect in the current item set.

2.5.6. Body-N effects for words in lexical decision

The last set of analyses was preformed on lexical decision latencies for words. The dependent and independent variables were identical to the reading aloud for words analyses.

2.5.6.1. LME model analysis

A model comparison showed a significant advantage of a model including all three-way interactions over one including all two-way interaction, χ^2 (4) = 14.20, p < 0.01, but no further improvement of a model including the four-way interaction, χ^2 (1) = 2.19, p > 0.1. The results of the body-N model including all three-way interactions are summarised in Table 5.

We found a facilitatory effect of frequency. There was a significant interaction between body-N and orthographic N, which we followed up by assessing the body-N slope at high (1 SD above the mean) and low (1 SD below the mean) orthographic N. This showed a facilitatory trend, but no significant body-N slope at high orthographic N words, t = -1.24, p > 0.2. At low orthographic N, the body-N slope was positive, indicating an inhibitory body-N effect, and reached significance, t = 1.99, p < 0.05.

Similarly, we followed up on the two-way interaction between body-N and frequency by assessing the slope at high and low frequency. At high frequency, we found a non-significant trend and a negative slope, suggesting a facilitatory body-N effect, t = -1.28, p > 0.2. At low frequency, we found a non-significant trend and a positive slope, suggesting an inhibitory body-N effect, t = 1.62, p > 0.1.

The third significant two-way interaction occurred between frequency and language. We therefore assessed the slope of frequency at each language. At English, the frequency slope was highly significant and showed a facilitatory effect of frequency, t = -10.27, p < 0.0001. At German, the slope of frequency was also significant and facilitatory, t = -3.84, p < 0.001. The interaction occurred because the slope was steeper for English (estimate = -0.32) than German (estimate = -0.07).

Table 5.

Output from the LME analysis for lexical decision of words, including body-N, Language, Orthographic N, frequency, and the interactions, and the main effect of onset complexity.

	Estimate	Std. error	t value	p value
Intercept	*-1.613*	*0.081*	*-19.989*	*0.001*
Main effects				
Body-N	< 0.001	0.001	0.299	0.765
Language	0.032	0.078	0.406	0.723
Orthographic N	-0.001	0.003	-0.281	0.779
Log frequency	-0.099	0.011	-9.175	< 0.001
Onset complexity	0.002	0.014	0.111	0.912
Two-way interactions				
Body-N x language	0.001	0.001	1.182	0.238
Body-N x orthographic N	-0.001	< 0.001	-2.074	0.039
Body-N x log frequency	-0.004	0.002	-2.384	0.018
Language x orthographic N	0.003	0.003	0.959	0.338
Language x log frequency	0.033	0.011	3.020	0.003
Orthographic N x log frequency	-0.003	0.004	-0.646	0.518
Three-way interactions				
Body-N x language x orthographic N	-0.001	0.000	-1.889	0.060
Body-N x language x log frequency	0.002	0.002	1.054	0.293
Body-N x orthographic N x log frequency	< 0.001	< 0.001	0.818	0.414
Language x orthographic N x log frequency	-0.011	0.004	-2.806	0.005

Finally, the LME model suggested that there are two three-way interactions: (1) The interaction between language, orthographic N and frequency was significant. We chose not to follow up on this, as it does not include our effect of interest (i.e., body-N). (2) The three-way interaction between body-N, language, and orthographic N was marginally significant. Examining the body-N slope at English high orthographic N, English low orthographic N, German high orthographic N and German low orthographic N showed that the body-N slope was significant at German for low orthographic N words, showing an inhibitory effect of body-N, $t = 2.53$, $p < 0.05$, all other $p > 0.3$. Numerically, the body-N slope remained negative (suggesting facilitation) for English, and was also negative for German at high orthographic N.

In summary, the LME analyses showed no main effect of body-N, but several interactions. These appear to be due to cross-overs, where the body-N is facilitatory (but non-significant) for high orthographic N words, but inhibitory (also non- significant) for low orthographic N words. This pattern appears to be driven by German: the body-N slope for English remains non-significant, but shows small facilitation.

2.5.6.2. Bayes Factor analysis

To mirror the LME analyses, we first constructed a set of models to assess the stability of the interactions. We constructed a set of BF models: one including only the main effect, another with the main effect and all two-way interactions, another model also containing all three-way interactions, and a full model which in addition contained the four-way interaction. The evidence for the two-way interaction compared to the main effects model was equivocal, BF = 0.79 (±2.68%), as was the evidence for the two-way compared to the three-way interaction model 0.38 (±2.9%). The main effect model, however, was supported over the three-way interaction model, BF = 3.38 (±1.61%) and over a four-way interaction model, BF = 7.43 (±1.56%). Therefore, the evidence suggests

that a main-effects only model performs significantly better than the models including the three- and four-way interactions, and numerically better than the two-way interactions model.

Excluding all interactions, we compared a model including body-N to one excluding it. Here, we found evidence against the model which included body-N, BF = 0.11 (±3.36%). Furthermore, we examined the theoretically important interactions between body-N and language, as well as body-N and frequency. Here, the evidence for the body-N by language interaction was 0.32 (±1.28%), suggesting that body-N does not interact with language, and the evidence for the body-N by frequency interaction was 0.87 (±1.26%), thus showing equivocal evidence for the presence of the frequency by body-N interaction, but being slightly in favour of the no-interaction model.

In addition, we compared a model which included the interaction of language and frequency, to one that included only main effects (but excluded the body-N effect, because we had show evidence against it). Here, the evidence was equivocal, BF = 2.32 (±3.46%).

Summary

As for the reading aloud word results, the analyses for lexical decisions of words seem to be characterised by higher-order interactions according to the LME analyses, although the BF analyses showed little support for any interactions. None of the analyses, however, showed any evidence for the presence of a body main effect, nor for an interaction with language. The LME analyses showed that the two- and three-way interactions which include body-N are driven by cross-overs, where the slope is positive (inhibitory) for low orthographic N words and low frequency words, and negative (facilitatory) for high orthographic N and high frequency words. A follow-up of the

marginally significant three-way interaction showed that the interactive pattern seemed to be driven by an inhibitory body-N effect for German low orthographic N words.

It is worth noting that the frequency by language interaction emerged in the LME analysis, and the BF gave weak evidence for this interaction (though it did not exceed the critical value of 3). As it also emerged in some analyses of words in reading aloud, this trend might be worth investigating in future research. In particular, the frequency effect is stronger for English than for German. If this effect is replicable, it might provide support for the Orthographic Depth Hypothesis (Katz & Frost, 1992): the Orthographic Depth Hypothesis proposes stronger involvement of the lexical route in a deep compared to a shallow orthography. Therefore, it would predict that we would find a stronger effect of a lexical marker, such as frequency, in English compared to German.

2.5.7. Discussion

In this discussion, we focus on those results that relate to our original questions, namely: (1) the stability of the main effect of body-N, especially across lexicality and task (2) interactions of body-N and frequency, and (3) interactions of body-N and language.

2.5.7.1. Main effects of body-N

The only condition which showed a stable effect of body-N was lexical decision for nonwords. Here, a higher body-N lead to longer latencies, meaning that high body-N nonwords are more difficult to reject than low body-N nonwords. The body-N effect seems to exist in addition to an inhibitory orthographic N effect (which is relatively consistently reported in the existing literature on orthographic N; for a review, see Andrews, 1997). This suggests that bodies reflect some aspect of the lexical system: a high body-N nonword appears to cause lexical activation of its body neighbours, and this lexical activation makes it more difficult to determine that it is a nonword.

In the other conditions, there was no trace of a main effect of body-N, suggesting that a body-N manipulation is insensitive to picking up reliance on body units during single-word and nonword reading aloud and in lexical decision for words. Any significant interactions that we found in the LME analyses seemed to be driven by cross-over patterns, which were furthermore not supported by the BF analyses.

We are not implying that bodies have no psychological reality, as this would be inconsistent with a growing body of research using different paradigms showing reliance on bodies. Therefore, the appropriate interpretation of the absent body-N effect in three out of four conditions is that the body-N manipulation is not a sensitive measure of reliance on bodies. As mentioned in the introduction, nonword reading studies that manipulate the existence versus non-existence of a body in real words (is *dake* easier to read than *daik*?) consistently show body effects (Andrews et al., 2005; Goswami et al., 2003; Rosson, 1985; Treiman et al., 1990), as do nonword reading studies, where the use of bodies would predict a different pronunciation compared to grapheme-phoneme correspondences, such as *dalk*, which can be read to rhyme with "talk" (if body-rime correspondences are used), or "talc" (if GPCs are used; Andrews & Scarratt, 1998; Brown & Deavers, 1999; Glushko, 1979; Schmalz et al., 2014, i.e., Paper 4).

Taking into account both previous body-existence effects and the findings of our body-N analyses, it might be suggested that body-related effects operate in a logistic fashion: a word with a known body is easier to process in a reading aloud task than a word with an unknown body. Once the body is established as a salient unit, it does no longer matter whether it occurs frequently, as additional exposure to the same body might not lead to further reinforcement of its saliency. This might partly explain the mixed results from the existing body-N studies: if a study contains bodies with very low body-N, or ones which occur only in low-frequency words, these may not be established in the

participants' cognitive systems, resulting in slower average processing of low- compared to matched high body-N items.

If this explanation holds true, we can make some predictions about individual differences in the size of the body-N effect, which might be followed up in future research. For example, we might expect a stronger body-N effect in children, as they may still be in the process of refining their lexical and sublexical systems, and low-frequency bodies may not yet be established (see Paper 3). Furthermore, we might expect individual differences in the degree to which skilled readers establish body representations, as opposed to relying on grapheme-phoneme correspondences (e.g., as a function of reading instruction; see Thompson, Connelly, Fletcher-Flinn, & Hodson, 2009).

2.5.7.2. Bodies across languages

Overall, we are confident in concluding that in the meta-analyses of nine studies that were included, the body-N effect does not differ across languages. The only condition where a body-N by language interaction emerged was the reading aloud of nonwords condition, and here we showed that this was due to a non-significant facilitatory trend for English and a non-significant inhibitory trend for German. As there are no bases for expecting an inhibitory body-N effect for reading aloud nonwords in German, let alone different directions of the body-N effect across languages, we attribute this pattern to a Type I error.

The absence of a main effect limits the strength of the conclusions we can draw about cross-linguistic differences in reliance on bodies: as the body-N effect is clearly not sensitive to picking up reliance on bodies in the first place, a lack of an interaction does not necessarily imply that reliance on bodies does not differ across orthographies. However, two key studies that have been used to back up the assumption of a cross-linguistic difference are based on the body-N effect (Ziegler et al., 2001; Ziegler, Perry,

Ma-Wyatt, et al., 2003). Therefore, our results at least call for a thorough scrutiny of this assumption, and for further studies using alternative marker effects of body processing (see Papers 1, 3, 4, & 5).

2.6. Published data: Can it be explained by participant-level differences?

Unfortunately, we had to exclude the published studies on body-N effects in skilled readers, as the trial-level data for these were not available. This may be argued to create a "reverse" file-drawer problem (Rosenthal, 1979), as both published studies showed significant facilitatory effects of body-N while the results of the unpublished studies were often null-results.

While our meta-analyses showed no body-N effects in reading aloud, Ziegler et al. (2001) report a facilitatory body-N effect for reading aloud of words and nonwords. Their body-N effect was stable for English readers, but only marginally significant for German readers. There was no interaction with lexicality. Ziegler and Perry (1998) report the results of a lexical decision task, and found a facilitatory body-N effect for words, but no effect for nonwords. This contrasts with our finding of an inhibitory body-N effect for nonwords in lexical decision, and no effect for words.

The discrepancy between our results and those reported by Ziegler and colleagues may be due to individual variability of the participants. We therefore compared the available data from the published studies to data for the same items from other datasets. The mean reaction times for each item (i.e., item-level data) from the Ziegler et al. (2001) study were taken from the appendix of Perry and Ziegler (2002). For the Ziegler and Perry (1998) study we took the averages for each condition (i.e., study-level data; p. B57, Ziegler & Perry, 1998).

We compared the Ziegler et al. (2001) data to data which we have collected using an identical item set (albeit with a few missing items). This allowed us to compare their

obtained item-level effects with ours. For the Ziegler and Perry (1998) data, we compared the overall condition means to those of the English Lexicon Project (ELP; Balota et al., 2007) and British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012) for the same items. As the effect of interest in the Ziegler and Perry (1998) is for words, we were able to find corresponding entries both in the ELP and BLP for all of the critical items.

2.6.1. Ziegler et al. (2001) data

We started with a re-analysis of the item-level data of the Ziegler et al. (2001) study. In contrast to the original analysis, we used body-N as a continuous predictor rather than a dichotomy of a high and low body-N condition. As described in the introduction, this removes the bias of a stronger body-N manipulation for their English compared to the German item set.

A linear model including body-N (as a continuous variable), lexicality, language, and their interactions as predictors, and RT as the outcome variable, showed a significant effect of lexicality, $t = -5.96$, $p < 0.0001$, with faster responses for words than nonwords. All other $p < 0.1$. Importantly, the facilitatory effect of body-N did not approach significance, $t = -1.53$, $p = 0.13$, and neither did the body-N by language interaction, $t = 1.34$, $p = 0.18$.

In a BF analysis, we compared a full model including body-N and all interactions to an identical model, excluding body-N and its interactions. We obtained evidence against the full model, BF = 0.02 (±4.14%). Comparing a main-effects model only to a main-effects model which excluded the effect of body-N gave equivocal evidence for H_1, BF = 0.44 (±3.73%). To assess the evidence for a body-N by language interaction, we compared a model including the interaction and main effects of body-N and language, as well as lexicality and its interaction with language, to an identical model which excluded

the interaction of body-N and language. Here, we again obtained equivocal evidence for H_1, BF = 0.73 (±8.92%).

Next, we turned to our own data, which we collected with the item set of Ziegler et al. (2001; see Table 2). In an identical analysis, we found a facilitatory main effect of body-N, t = -2.79, $p < 0.01$, a main effect of lexicality, t = -6.91, $p < 0.0001$, with faster responses to words than nonwords, and a body-N by language interaction, t = 2.11, $p < 0.05$, with a stronger body-N effect in English than in German. Post-hoc tests for each language individually showed that in English, the body-N effect was significant and facilitatory, t = -2.74, $p < 0.01$ (as was the effect of lexicality, t = -6.79, $p < 0.0001$). For German, there was no body-N effect, $t < 1$ (but a main effect of lexicality, t = -5.26, $p < 0.0001$).

In summary, in a re-analysis of the RTs of Ziegler et al. (2001) as reported by Perry and Ziegler (2002), we find equivocal evidence for a body-N effect, and an interaction with language. Thus, it appears that neither the body-N effect, nor its interaction with language, remain stable when body-N is treated as a continuous variable rather than a dichotomy. This suggests it is possible that the language-by-body-N interaction reported by Ziegler et al. (2001) is a result of their stronger body-N manipulation for English than German.

In a re-analysis of our own data using the same items, we find the pattern originally reported by Ziegler et al., (2001): a significant body-N effect for English, and none at all for German. This contradicts the LME analyses for the same data that we present in Table 2: Here, we had shown evidence against a body-N effect for German nonwords, BF = 0.12, English words, BF = 0.20, and English nonwords, BF = 0.28. The evidence for a body-N effect in German words was equivocal, BF = 0.73.

The major difference between the analyses reported in this section and those in Table 2 is that here we did not include subject as a random factor, as we had averaged the RTs for each item across subjects in order to directly compare our data to the item-level data reported by Perry and Ziegler (2002). One possibility is that body-N effects are highly variable across participants, and that when such effects are found they may be driven by only a few subjects. For future research, a valuable approach to test this hypothesis would be to conduct experiments to isolate additional factors that predict reliance on body units, and the extent to which orthographic depth plays a role.

2.6.2. Ziegler and Perry (1998) data

Neither the trial-level not the item-level data for the Ziegler and Perry (1998) lexical decision study are available. In that study, the authors report a facilitatory effect of body-N for words, but not for nonwords. The average reaction times for the high- and low body-N conditions were 625 ms and 657 ms, respectively, suggesting a "strong" (F_1 = 20.82, F_2 = 4.96; pp. B57-B58) facilitatory body-N effect.

This conflicts with the analyses that we presented above, where we showed an inhibitory body-N effect for nonwords in lexical decision, and no effect of body-N in lexical decision for words. As seen in Table 2, the study of Marcus Taft provided evidence against an effect of body-N on lexical decision latencies for words, BF = 0.10. Of our two lexical decision studies for words, one provided equivocal evidence, BF = 1.55, and the other provided evidence against the body-N effect, BF = 0.18. The discrepancy in the nonword lexical decision results could be attributable to low power in the Ziegler and Perry study: due to their care in matching for potential confounds, their item set was reduced to 20 items per condition. Explaining the discrepancy in the lexical decision word results is less straightforward.

To examine whether the body-N effect for words reported by Ziegler and Perry is stable (1998), we obtained the by-item lexical decision means for the high and low body-N words that were used by Ziegler and Perry (1998) from the ELP (Balota et al., 2007) and the BLP (Keuleers et al., 2012). In the ELP, the average latencies for the high and low body-N conditions were 652.03 ms (SD = 86.62) and 665.68 ms (SD = 71.83), respectively. A *t*-test showed that this 12 ms advantage for high body-N words was not significant, $p > 0.5$. In the BLP, the average latencies for the high and low conditions were 582.67 ms (SD = 59.60) and 572.52 ms (SD = 40.21), respectively. This 11 ms difference in the unexpected direction was also not significant, $p > 0.5$.

2.6.3. Discussion

The analyses comparing the published studies to data using the same items but different participants are interesting, because the discrepancy in the results suggests that the variation in the size of the body-N effect may be attributable to participant-level factors rather than item-level factors. For the Ziegler et al. (2001) study, their body-N effect and the interaction with language becomes less robust when we treat body-N as a continuous variable, but we find a significant interaction in our data, which used an (almost) identical dataset.

Ziegler and Perry (1998) found a strong facilitatory body-N effect while, for the same words, the ELP data show a non-significant facilitatory trend and the BLP data show a non-significant inhibitory trend. This supports our tentative conclusion from the previous section, that sensitivity to body-N may differ across individuals, which may account for some of the discrepant findings reported throughout this paper.

2.7. General discussion

In the current paper, we attempted to draw together an accumulating amount of research results on the body-N effect. We know of only two published studies that have

reported a significant body-N effect in skilled adult readers (Ziegler & Perry, 1998; Ziegler et al., 2001); findings from our own lab and an unpublished study from another lab (M. Taft, personal communication, 2014) provide inconclusive results. Taking a closer look at the unpublished null-results is important, because a focus on published positive results may be subject to the file-drawer problem, where null-results remain unpublished and the scientific community relies on a small percentage of type-I error results (Rosenthal, 1979).

The first conclusion from our meta-analysis is that the body-N effect, where items that have many body neighbours are compared to items that have fewer body neighbours, is an unreliable measure of reliance on body units. Although studies using different paradigms have shown that participants rely on these units, we find no trace of a main effect of body-N in most conditions. This has methodological implications for future studies: conducting an experiment with a body-N manipulation may not be the best approach to exploring reliance on body units. More reliable measures appear to be nonword reading paradigms, where the existence or otherwise of the body is manipulated, such as comparing reading latencies to the nonwords *dake* and *daik* (Andrews et al., 2005; Goswami, Gombert, & de Barrera, 1998; Goswami, Porpodas, & Wheelwright, 1997; Goswami et al., 2003; Rosson, 1985; Treiman et al., 1990), or where the pronunciation differs depending on whether participants use grapheme-phoneme correspondences or body-rime correspondences, for example, whether *dalk* is pronounced as "/dælk/" or "/do:k/" (Andrews & Scarratt, 1998; Brown & Deavers, 1999; Glushko, 1979; Schmalz et al., 2014; Thompson et al., 2009). A drawback of nonword reading experiments, however, is that it is impossible to apply the same procedure to words, meaning that theoretically important interactions with lexicality and frequency cannot be assessed.

A question that remains to be answered is why some previous studies have found a significant effect of body-N. Although these could be false positives, it is also possible that they reflect a real effect of body-N in some participants. The results of our analyses are consistent with the notion that a body-N effect is present in some individuals, but not others. We cannot determine what factors underlie these individual differences - this remains a question for future research.

According to the psycholinguistic grain size theory, the inconsistency of the grapheme-phoneme correspondences is one of the pressures which pushes for reliance on body-rime correspondences. This is backed up by studies of the body-N effect (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003), which showed a stronger body-N effect in English compared to German. Due to methodological issues with their item set, our second aim was to assess the stability of the interaction between body-N and language. The meta-analyses showed little evidence for a body-N by language interaction. In a re-analysis of the item-level data of Ziegler et al. (2001), the evidence for a body-N effect or an interaction with language became much less robust when we used body-N as a continuous variable. Our own replication of the experiment of Ziegler et al. (2001), using their items, did provide evidence for the presence of an interaction between language and body-N. However, when we used subjects as a random factor (cf. Table 2 and the Appendix), we mostly obtained evidence against a body-N effect. We therefore conclude that the interaction found in our item-level analysis occurs because it averages out individual differences, where only a subset of English speakers show a body-N effect.

As discussed in the Introduction, using body-N as a continuous variable rather than a dichotomy reduces the biases associated with a stronger body-N manipulation for English than German in their item set. Therefore, taken together, our findings suggest that the results originally reported by Ziegler et al. (2001 & 2003) may be due to a stronger

body-N manipulation for English than German, rather than a difference that is driven by orthographic depth.

In addition, the results presented throughout this study suggest that there might be individual variability, where some unknown factor causes some participants to show sensitivity to body-N, but not others. At this stage, we can only speculate what this factor might be. A promising possibility, which would explain that showing a body-N effect seems more common for English than German-speaking participants, might be reading instruction. In German, due to the close correspondence between print and speech, it is customary to teach children to read via phonics instruction. In English, conversely, a mixture of phonics and whole-word instruction is common (Landerl, 2000). A recent study has used a nonword reading paradigm to show that reliance on bodies is more common for adult participants who had received whole-word reading instruction as children, while participants who had received phonics instructions tended to give nonword responses that were in line with grapheme-phoneme correspondences (Thompson et al., 2009). Therefore, future research is needed to establish whether cross-linguistic difference in the reliance on "large" units persist once effects of reading instructions are controlled for.

The third aim of the study was to isolate the locus of a body-N effect. In the introduction, we outlined a set of predictions that would allow us to determine whether bodies are processed by a sublexical route (M. Coltheart, 2012; M. Coltheart et al., 2001; Perry et al., 2007), or whether they are more likely to represent lexical activation from body neighbours (Forster & Taft, 1994). Although we cannot draw any strong conclusions due to an overall absence of a main effect of body-N for most conditions, our finding of an inhibitory body-N effect for nonwords in lexical decision suggests that there is a lexical locus of the body-N effect. This can be explained by an interactive activation

account, as this finding suggests that a high body-N nonword causes some activation of lexical nodes. This activation, in turn, makes it harder for the participant to reject a nonword. Similar findings have been reported for orthographic neighbours (reviewed in Andrews, 1997). We found an effect of both body-N and orthographic N, which suggests, on the one hand, that body-N effects exist even when controlling for orthographic N, and on the other hand that orthographic N effects exist even after controlling for body-N. This provides a new benchmark for computational models of reading.

In conclusion, the results of our meta-analysis challenge previous findings of a body-N effect. We find little evidence for a stable body-N effect. These outcomes call for a re-examination of the claim that reliance on body units differs across orthographies, as this conclusion is based on findings of interactions between body-N and language (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003), which we have shown to be unstable, and possibly driven by factors other than orthographic depth. Establishing what exact factors influence reliance on bodies, and whether orthographic depth makes any contribution, remains a question for further research. Finally, future work, using different manipulations, is needed to determine the exact mechanisms via which bodies and other larger-than-grapheme units operate within the cognitive system, as this has important implications for models of reading aloud.

Paper 3: Body-N Effects across Reading Acquisition

Abstract

A debate in the literature on reading relates to the specific mechanisms that compute the pronunciation of unfamiliar words. In relation to reading acquisition, reliance on "large" orthographic units (e.g., the body *-alk* to read the nonword *dalk*) has been proposed to differ as a function of age and characteristics of the orthography. Here, we follow up on these claims, using the body-N effect to measure reliance on large units. In Experiment 1, we find a stable body-N effect for nonwords, but not words, in German children in grades 2-4. The lexicality interaction suggests that large bodies operate via a mechanism that is independent of whole-word processing. The stability across age shows that by Grade 2, the non-lexical mechanisms underlying reading in German are already well-refined. In Experiment 2, we test bilingual English/German children to provide a strong test of the claim that reliance on bodies differs as a function of orthographic characteristics. When the same children read German compared to English items, the size of the body-N effect did not differ, suggesting that previous reports of a body-N by language interaction may be driven by uncontrolled differences on the item or participant level.

3.1 Reliance on body units: Body-N effects across reading acquisition

Reading unfamiliar words involves relying on knowledge about sublexical orthographic units, such as letters, and how they map onto sounds. A strongly debated question in the literature on reading acquisition relates to the types of orthographic units that are used by children as they learn to decode new words (Brown & Deavers, 1999; Goswami, 2002; Goswami & Bryant, 1990; Hulme et al., 2002; Nation, Allen, & Hulme, 2001; Ziegler & Goswami, 2005). Previous research suggests that the units of choice depend, among other factors, on the age of the children (Goswami, 1993), and the nature of the orthography in which a child learns to read (Goswami et al., 1998; Goswami et al., 1997; Ziegler, Perry, Ma-Wyatt, et al., 2003).

The aims of the current study were twofold: First, we assessed the developmental trajectory of the reliance on body units during word and nonword reading in German. This was designed to provide insights into how and why reliance on bodies develops in a "shallow" orthography, such as German. Second, we followed up on an important finding from a previous study (Ziegler, Perry, Ma-Wyatt, et al., 2003), which suggests that the nature of sublexical units differs across orthographies varying in orthographic depth (Ziegler & Goswami, 2005).

3.1.1. What are bodies, and how are they processed?

Sublexical decoding, or deriving the pronunciation of a letter string via a non-lexical assembly procedure, is crucial for reading acquisition, as it is considered to be an essential mechanism for establishing word knowledge (Share, 1995). Therefore, it is important to understand the mechanisms via which it operates, and how these develop. One focus of attention in relation to this question has been on body units, as evidence suggests that these orthographic units have strong psychological reality (for reviews, see Goswami, 1999; Goswami & Bryant, 1990; Ziegler & Goswami, 2005). Bodies consist of

the vowel and coda letters of a syllable, such as *-iend* in the word *friend*. As sublexical units, bodies are often contrasted with graphemes (e.g., Andrews, 1982; Cortese & Simpson, 2000; Jared, 2002), which are letters or letter clusters that are used to represent a single phoneme, such as *t*, *th*, or *ough* (M. Coltheart et al., 2001). Previous work has focused predominantly on English, however, evidence also suggests that body units have psychological reality, to some extent, in other orthographies, such as French, Spanish (Goswami et al., 1998), and German (Goswami et al., 2003; Ziegler, Perry, Ma-Wyatt, et al., 2003).

In the current study, we use the body-N effect as a marker of reliance on body units (Ziegler & Perry, 1998; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). The body-N size of a given letter string is measured as the number of words that have the same body. For example, the nonword *lat* has 16 body neighbours, including the words *at*, *hat*, and *chat*, whereas the nonword *lazz* only has one body neighbour, *jazz* (Ziegler et al., 1997). If items with many bodies are read aloud faster than items with fewer body neighbours, this suggests that information about bodies is in some way drawn on when performing the task. Although this effect has been demonstrated in previous studies (Ziegler & Perry, 1998; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003), it is unclear what exactly the body-N effect measures, and whether it reflects a lexical analogy strategy, or a purely sublexical process.

In the developmental literature, bodies have been considered in terms of a lexical analogy mechanism. In a set of experiments, Goswami (1991a, 1991b) presented children with words that they did not know and with a clue word which overlapped either with the body (e.g., *b<u>eak</u>* - *p<u>eak</u>*), or the antibody (e.g., *<u>beak</u>* - *<u>bean</u>*). The body analogy condition was particularly helpful to children in deciphering the unfamiliar word (Goswami, 1991). This view on how bodies are used suggests the direct involvement of the orthographic

lexicon. Combining it with the adult literature on interactive activation in lexical processing (e.g., Taft, 1991), a cognitive explanation of how body-N effects occur could be as follows: when a word is encountered, it activates the corresponding entry in the orthographic lexicon, but also that of words that are close in the orthographic space (Forster & Taft, 1994). Activation between the orthographic nodes can facilitate the recognition of the target, leading to faster recognition of a high body-N compared to a low body-N word. As whole-word activation also feeds back into sublexical units, facilitation for nonwords associated with a high body-N could also be expected. If this is the case, then a study which finds an effect of body-N while keeping other orthographic similarity measures constant gives us valuable information about the structure of the orthographic lexicon: it suggests a hierarchical organisation of the orthographic space, where bodies are more important in retrieving a lexical representation than other units, such as antibodies (Forster & Taft, 1994; Ziegler & Perry, 1998).

An alternative view is one where bodies are processed as sublexical units, in an analogous fashion to graphemes (Patterson & Morton, 1985). Here, there is no direct connection between lexical entries and bodies. Rather, the sublexical route might parse a letter string simultaneously into grapheme and body units, which would be mapped to their phonological equivalents (phonemes and rimes) and re-assembled into a whole-word pronunciation. The speed with which the rime is accessed might depend on the frequency of the body, if we assume that sublexical units that occur often have a lower activation threshold.

Thus, although the reliance on body units in children has been studied extensively (Brown & Deavers, 1999; Goswami, 1991; Goswami et al., 2003; Ziegler, Perry, Ma-Wyatt, et al., 2003), it is unclear what reliance on body units means, as the two scenarios described above provide different explanations about the cognitive mechanisms

underlying the body-N effect. To address this question, we manipulated lexicality as well as body-N. If the body-N effect reflects activation between lexical nodes, the effect may be stronger for words than for nonwords. If the body-N effect reflects the activation of sublexical body units, we would expect the effect to be stronger for nonwords than for words, because the lexical influence involved in word naming might be achieved prior to any sublexical influence.

A previous study has assessed body-N effects in German and English-speaking children who were either dyslexic or normal readers (Ziegler, Perry, Ma-Wyatt, et al., 2003). The children in this study were asked to read aloud both words and nonwords with a high or low body-N. There was no interaction between body-N and lexicality. In this analysis, however, lexicality and its interactions were not a main focus, and the complexity of the analyses may have reduced the power necessary to detect any possible interactions between lexicality and body-N. Furthermore, their lexicality manipulation may have been weakened by the presence of a number of low frequency words in the item set. As finding an interaction between lexicality and body-N has theoretical implications, our first aim was to replicate their finding in a study more directly focussed on lexicality effects.

3.1.2. Reliance on bodies across age

The second aim of the current study was to explore the developmental trajectory of reliance on bodies in German. Here, we aimed to addresses an ongoing debate in the literature on reading acquisition, relating to the size of the sublexical reading units used by young readers. Specifically, we sought to assess how reliance on bodies changes across development, and the extent to which findings from English are generalisable to orthographies such as German, which are more representative of an average European orthography (Seymour et al., 2003; Share, 2008). Although some evidence exists for the

use of bodies in orthographies other than English (Goswami et al., 1998; Goswami et al., 2003; Ziegler, Perry, Ma-Wyatt, et al., 2003), the majority of studies on the reliance on bodies involve English-speaking children.

In the context of the large-versus-small units debate, letter or grapheme units are often referred to as "small" units, and body units are called "large" units. Large-units-first theorists propose that in the beginning of reading acquisition, children rely on large units (Goswami, 1991, 1993, 1998, 2002; Goswami & Bryant, 1990). This is hypothesised to occur because the phonological awareness associated with small units does not develop until the onset of reading instruction (see Castles & Coltheart, 2004 for a review), which prohibits the use of small units. As a result, larger units are more prominent in allowing children to make analogies to words with similar spelling patterns. According to this view, learning to read involves developing increasingly refined reading analogy mechanisms for unfamiliar words; due to the emergence of phoneme awareness, the reliance on small units becomes more important in older readers (Goswami, 1993).

Small-units-first theorists argue that phonemic awareness is a better predictor of reading ability than large-unit phonological awareness (Duncan et al., 1997; Hulme, Bowyer-Crane, Carroll, Duff, & Snowling, 2012; Hulme et al., 2002). This is taken to indicate that small units are more important in early reading than large units. According to this view, learning to decode involves the interaction of various processes: in the beginning, children use their knowledge of basic letter-sound correspondences to decode unfamiliar words. As words become familiar and the mental orthographic lexicon grows, whole-word knowledge can be used to derive more subtle regularities between print and speech. Children thus learn about units which were not explicitly taught via bootstrapping from lexical knowledge (Ziegler et al., 2014). As a result, simpler, smaller

correspondences are acquired first, and with reading experience children also learn the statistical regularities underlying bodies and other statistically salient units.

The small-units-first and large-units-first theories of learning to read make clear predictions regarding the body-N effect across age. Taking the body-N effect as a marker effect of large-unit processing, we expect that the strength of the effect would diminish with age according to the large-units-first hypothesis; the small-units-first hypothesis predicts that the strength of the effect should increase with age.

A previous study has shown a body-N effect for both word and nonword reading in English-speaking and German children (Ziegler, Perry, Ma-Wyatt, et al., 2003). This suggests that body-N effects can pick up reliance on body units in children. An outstanding question is the developmental trajectory of the effects: Ziegler et al. (2003) compared dyslexic children to matched controls, and did not set out to examine differences across grades. Therefore, the theoretically important question of changes of the body-N effect across age remains unanswered.

3.1.3. Body-N effects across orthographies

Previous research has proposed that the types of orthographic units are dependent on the orthography (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). According to the Psycholinguistic Grain Size Theory (PGST; Ziegler & Goswami, 2005), orthographies with unreliable correspondences (so-called "deep" orthographies) force readers to rely on larger units to reduce the ambiguity of the pronunciation. This has been backed up by studies of the body-N effect: Ziegler et al. (2003) have shown that this effect is stronger in English than German children (German orthography is considered "shallow", because the grapheme-phoneme correspondences are predominantly reliable, whereas English is widely cited as an example of a deep orthography).

There are two aspects of the study of Ziegler et al. (2003) that limit the strength of their conclusion that reliance on body units differs between English and German. In the current study, we aimed to address these. Firstly, as with all cross-linguistic studies, it is very difficult to match for a wide range of factors that may differ across countries. Whole-word instruction methods are more common in English than in German-speaking countries (Landerl, 2000). It has been previously shown that phonics reading instruction biases readers towards reliance on grapheme-phoneme correspondences during nonword reading, whereas whole-word reading instruction pushes for body-consistent nonword pronunciations (Thompson et al., 2009). Therefore, failing to control for reading instruction in a cross-linguistic design provides a plausible alternative explanation for increased reliance on bodies in English compared to German.

Additionally, when matching children across orthographies, decisions need to be made about what measures to match them on. A similar problem has been described reading-age matched designs in dyslexia research (Jackson & Coltheart, 2001): for example, when the two samples are matched on single-word reading fluency, they may still differ in other measures, such as nonword reading accuracy or text comprehension. In an across-language design, such differences may be particularly critical due to different age of reading instruction onset and teaching methods. In Experiment 2, we circumvent these problems by using a within-participant design: we report data with children from a bilingual German/English school in Australia, who read both English and German item sets manipulated by body-N.

The second limitation of the Ziegler et al. (2003) study relates to the use of cognates: an advantage of the cognate design is that the items are mostly identical in terms of their orthography and meaning across languages, which allows for control over characteristics such as frequency, imageability, and length. A drawback of this design is

that cognates tend to differ systematically across orthographies on lexical and sublexical variables as a function of broader linguistic characteristics. For example, the items used by Ziegler et al. (2003) were not matched for orthographic N or body-N across languages. This resulted in a stronger manipulation of body-N for English than German: In English, the average body-N values were 15.33 (SD = 6.00), and 4.66 (SD = 3.75) for the high and low body-N conditions respectively, resulting in a manipulation of 10.67. For German, the high and low body-N conditions had, on average, 10.28 (SD = 4.35) and 4.29 (SD = 3.75) body neighbours, resulting in a manipulation of 7.99. Thus, it is unclear whether the language by body-N interaction can be traced back to the influence of orthographic depth, or whether it is a result of the density confound. To address this issue, we designed an item set which was matched as closely as possible across languages for item characteristics that are known to affect reading latencies, such as length, frequency, and orthographic N, while ensuring that the strength of the body-N manipulation did not differ across orthographies.

3.1.4. Aims

The experiments in the current study aimed to broadly address the questions of whether bodies are processed within the lexical or sublexical system, and how age and the characteristics of the orthography influence reliance on this orthographic unit. In Experiment 1, we explored the developmental trajectory of the body-N effect in German. Small-units-first theories predict that the size of the body-N effect should increase with age, whereas large-units-first theories predict that it should decrease. Furthermore, exploring interactions of body-N and lexicality will allow us to determine the mechanisms underlying the body-N effect.

In Experiment 1, we aimed to examine the developmental trajectory of reliance on body units, using the body-N effect. We applied the general procedure used by Ziegler et

al. (2003) to German children in grades 2-4. In these grades, children have mostly acquired basic decoding skills, and are in the process of building up a sound mental lexicon and refining their sublexical knowledge. Tracking the reliance on body units in this age range can show whether the fine-tuning of sublexical knowledge leads to increased reliance on small units, or whether it allows the children to learn correspondences associated with body units as a function of reading experience.

In Experiment 2, we followed up on the earlier finding of a body-N by language interaction in English- and German-speaking children (Ziegler, Perry, Ma-Wyatt, et al., 2003). We controlled for participant-level variability by using a within-subject design, and for item-level confounds related to the lexical density of the two orthographies by matching our items on body-N and orthographic N in English and German.

3.2. Experiment 1: Body-N Effects as a function of age

3.2.1. Methods

3.2.1.1. Participants

The participants were school children in Potsdam, Germany: 24 children in Grade 2, 19 children in Grade 3, and 22 children in Grade 4. Testing took place at Potsdam University, in a quiet room, between April and June, in the second half of the school year. In addition to the body-N items, the children were also tested on their sight-word reading ability with the 1-minute reading test of the Salzburger Lese- und Rechtschreibtest (SLRT II; Moll & Landerl, 2010). This test has been standardised, allowing us to classify the children's sight word reading ability relative to their peers. As seen in Table 1, the group scores were well within the average range as percentile scores between 16 and 84 are considered to be average. The participants' ages and performance on the sight word reading test are listed in Table 1.

Table 1.

Participant characteristics of Experiment 1; mean (SD)

School grade	Age (y;m)	SLRT raw score	SLRT percentile
2	8;0 (0;5)	46.3 (17.3)	50.6 (28.4)
3	9;0 (0;7)	73.0 (20.9)	62.8 (29.0)
4	10;0 (0;5)	76.3 (20.7)	59.7 (28.9)

3.2.1.2 Items

We used 90 words and 90 nonwords which differed in their body-N size (high, low). These are listed in the Appendix; the item characteristics are presented in Table 2 (for comparison, with the descriptive statistics of the matched English items, which we used in Experiment 2). Orthographic N and frequency values were taken from WordGen (Duyck et al., 2004), and the body-N values from type body-N counts by Ziegler and colleagues (Ziegler et al., 1997, for English; Ziegler, personal communication, 2012, for German).

Table 2.

Item characteristics of the German and English items, matched across body-N conditions and language; mean (SD).

	German				English			
	Words		Nonwords		Words		Nonwords	
	High BN	Low BN	High BN	Low BN	High BN	Low BN	High BN	Low BN
Body-N	13.69 (2.55)	2.00 (0.83)	13.36 (2.86)	2.20 (1.07)	13.31 (1.90)	2.18 (0.68)	13.38 (1.95)	2.18 (0.68)
Number of letters	4.51 (0.84)	4.13 (0.59)	4.53 (0.73)	4.09 (0.47)	4.47 (0.69)	4.27 (0.69)	4.53 (0.59)	4.31 (0.60)
Orthographic N	3.84 (2.82)	3.27 (2.20)	4.33 (3.61)	3.62 (1.87)	4.47 (1.62)	3.98 (2.15)	4.27 (3.17)	4.22 (2.75)
Log Frequency	1.56 (0.92)	1.45 (0.64)	0 (0)	0 (0)	1.66 (0.53)	1.63 (0.44)	0 (0)	0 (0)

3.2.1.3. Procedure

The experiment was controlled with the program DMDX (Forster & Forster, 2003). Each item was presented for 2500 ms or until the voice key was triggered. The children were instructed to read aloud the words or nonwords on the screen as quickly and accurately as possible. Words and nonwords were presented in separate blocks, the order of which was counterbalanced across participants. Within the blocks, the order of items was randomised.

3.2.2. Results

The data were scored offline with the program CheckVocal (Protopapas, 2007) as correct, incorrect, or no response. The RTs were readjusted to remove any potential biases associated with first phonemes or failures with the voice key trigger. Prior to conducting any analyses, we removed all non-responses (3.12% of all data). We further analysed the data with Linear Mixed Effect (LME) models, which allow us to use the by-item and by-subject variance as random factors, and thereby summarise the main effects and interactions in a single statistic (Baayen, 2008; Baayen et al., 2008). For accuracy, a binomial LME gives a *z*-value and its significance for each effect and interaction. For RTs, the LME gives a *t*-value for each main effect and interaction. Table 3 summarises the accuracy and RTs for all conditions.

We assessed the main effects and interactions of Grade (2, 3, or 4), lexicality (word, nonword) and body-N (high, low). In accuracy, the three-way interaction between grade, lexicality, and body-N reached significance, $z = 2.27$, $p < 0.03$. In addition, there was a significant main effect of grade, $z = 2.81$, $p < 0.01$, reflecting that accuracy increased with grade. The effect of lexicality (higher accuracy for words than nonwords) approached significance, $z = 1.75$, $p = 0.08$. All other $p > 0.15$. As the three-way

interaction for accuracy was significant, we performed separate post-hoc analyses for nonwords and words.

Table 3.

Body N effects for words and nonwords as a function of Grade; mean (SD).

	Reaction time (ms)				Accuracy (%)			
	Words		Nonwords		Words		Nonwords	
	High BN	Low BN	High BN	Low BN	High BN	Low BN	High BN	Low BN
Year 2	886.2	860.5	1038.9	1065.9	89.58	87.47	82.85	74.26
	(94.9)	(74.8)	(99.28)	(119.8)	(11.71)	(13.40)	(14.51)	(16.26)
Year 3	816.7	787.3	959.1	954.3	94.74	95.13	88.74	79.23
	(91.2)	(76.3)	(97.7)	(114.6)	(7.23)	(7.95)	(9.87)	(14.06)
Year 4	751.9	735.7	924.5	934.8	94.62	96.00	90.15	82.85
	(81.9)	(67.7)	(79.8)	(113.1)	(8.34)	(8.68)	(10.03)	(12.60)

As seen in Table 3, numerically, the body-N effect for nonwords does not appear to change to a great extent. We performed a post-hoc LME with grade and body-N as independent variables, in order to confirm that the body-N effect was stable across grades. This analysis showed a main effect of grade, $z = 2.82$, $p < 0.005$. The effect of body-N, and the interaction of body-N and grade, were not significant, $p > 0.1$. The interaction is theoretically important, as it is a measure of the extent to which the body-N effect is stable across the grades that we tested. A non-significant *p*-value cannot be interpreted, as it does not distinguish between the possibility that there is no effect, and the possibility that the manipulation is insensitive to picking up an effect. Therefore, we followed up with a Bayes Factor (BF) analysis (Rouder et al., 2009). A BF quantifies the strength of evidence for a particular model against an alternative model. BF values exceeding 3 are considered provide "some evidence" for the model that is tested; values exceeding 10 provide "strong evidence", and values greater than 30 provide "very strong

evidence" (Rouder et al., 2009). Values between 0.3 and 3 are considered to provide equivocal evidence for the model that is being tested and the model that it is being tested against, meaning that the data manipulation is insensitive, and the data cannot be used to distinguish between the two possibilities.

We used the R package "BayesFactor" (Morey & Rouder, 2014) to calculate the BF values for the *absence* of the grade by body-N interaction for nonwords only. We contrasted a model which included only the main effects of body-N and grade to one which contained the two main effects and also the interaction between the two factors. Item and subject were included as random factors. The BF value indicating the strength of evidence for the model *excluding* the interaction was 20.13 (±3.44%), providing strong evidence for the null hypothesis, that the size of the body-N effect does not change across the grades for nonwords.

In the LME analysis, the body-N effect for nonwords only was not significant. As this may compromise our conclusion of a stable body-N effect, we also performed a BF analysis, comparing a model excluding the body-N effect to one which included it, using the nonword data, to address the possibility that the LME result represents a false negative. The BF for the model which excluded the body-N effect was 0.08 (±2.28%), thus supporting H_1, over H_0, and suggesting that body-N has an effect on reading accuracy for nonwords, despite the non-significant *p*-value.

For words only, an LME analysis indeed showed a significant effect of grade, $z = 2.60$, $p < 0.01$, a marginally significant effect of body-N, $z = 1.75$, $p = 0.08$, and a significant interaction between grade and body-N, $z = 2.16$, $p < 0.05$, reflecting a facilitatory body-N effect in older grades and an inhibitory effect in younger grades. As this raises interpretational issues, we followed up with additional tests to assess whether in each grade, there was a significant body-N effect for words. For Grade 2, the

facilitatory body-N effect was not significant, $z < 1$, $p > 0.4$, and BF for $H_1 = 0.13$ (±2.39%). For Grade 3, $z < 1$, $p > 0.6$, and BF for $H_1 = 0.10$ (±1.65%). For Grade 4, $z = 1.14$, $p > 0.2$, and BF for $H_1 = 0.13$ (±1.05%). As BF values smaller than 0.3 provide evidence against the hypothesis that is tested, this suggests that there is no real body-N effect for words in either grade.

We conducted the RT analyses to assess the main effects of grade (2, 3, and 4), lexicality (words, nonwords) and body-N (high, low). The dependent variable was inverse RTs. We excluded all data points with inverse RT > 3 (i.e., < ca. 333 ms; 0.7% of all correct responses). A Q-Q plot indicated that the trimmed data approximated a normal distribution. The analyses showed a main effect of lexicality, $t = 5.28$, $p < 0.0001$, and a main effect of grade, $t = 2.89$, $p < 0.01$. All other $t < 1.4$. Neither the body-N effect, not its interactions approached significance, all $p > 0.6$.

3.2.3. Discussion

In the first experiment, we measured the body-N effect in German children in Grades 2-4. Overall, we found a body-N effect in accuracy, but not RTs. This is in contrast to the findings of Ziegler et al. (2003), who found the effect in accuracy and in RTs. The discrepancy between the two studies is not may simply be due to the fact that we used a more stringent marking criterion: as all our nonwords had regular and consistent bodies, we considered a response to be correct only when it rhymed with its base word. Ziegler et al. (2003) used a lenient marking criterion, where every plausible response was scored as correct. More importantly, we found, in accuracy, a significant three-way interaction between Grade, body-N and lexicality. Post-hoc tests showed that body-N and grade interacted for words, but not for nonwords.

The overall main effect of body-N (in accuracy) suggests that bodies have a psychological reality in German developing readers. There are two reasons why this is an

interesting finding. Firstly, German is considered to be a shallow orthography (Borgwaldt et al., 2005; Seymour et al., 2003). This means that knowledge of grapheme-phoneme correspondences should be mostly sufficient to achieve high accuracy. Secondly, body-rime correspondences are never taught explicitly in German schools. Therefore, the results suggest that the saliency of body units emerges in German children without strong pressure from orthographic inconsistency, or explicit reading instruction.

The lack of interaction between grade and body-N for *nonwords* suggests that for the three grades that were tested, the nature of sublexical processing is stable. This finding cannot be used to support either the small-units-first or the large-units-first theories of learning to read. Instead, it might suggest that German children already possess relatively refined sublexical systems by the end of grade 2, at least for the bodies which were used in the current experiment. Future research may further address this issue by using a body-N manipulation with younger children, or using items with lower body-N values, as for these may take longer to establish sound body representations.

For words, follow-up tests showed that the interaction between body-N and grade reflects different directional trends of the body-N effect in Grade 2 (facilitatory) compared to Grade 4 (inhibitory). In neither age group was the body-N effect significant. As the body-N effects, in the individual age groups, appear to be absent, it is unclear whether this interaction can be meaningfully interpreted. We know of no theoretical framework that would predict a negative body-N effect, when orthographic N is controlled for.

There are two possible explanations: (1) The trends for positive and negative body-N effects in the two age groups represent random noise, therefore the interaction is meaningless. (2) It might be possible that there are two loci for the body-N effect. One might produce inhibition, and the other might produce facilitation. For example, the

functioning of the sublexical route might create facilitation associated with a high body-N, and the lexical route may create inhibition associated with the effect. As the orthographic lexicons of the older children are more numerous and contain stronger connections, this inhibitory counter-acting of the sublexical facilitatory effect may be stronger for Grade 4 than Grade 2 children - which would produce exactly the pattern that we observe. As this explanation does not fall out of any model or theory that we know of, we do not discuss this further, but merely state it as a logical possibility.

3.3. Experiment 2: Body-N Effects in Bilingual Children

In Experiment 2, we tested German/English bilingual children on the same set of items as in Experiment 1, and on a matched English item-set. Here, we performed across-language analyses. Assessing the body-N effect across languages is theoretically important, as the PGST predicts differences in the reliance on body units associated with orthographic depth (Ziegler & Goswami, 2005). To our knowledge, our study is the first to use bilingual children to test a cross-linguistic theory of reading. Previous studies have compared monolingual children in their respective language (Bruck, Genesee, & Caravolas, 1997; Caravolas & Bruck, 1993; Caravolas et al., 2012; Caravolas, Lervag, Defior, Malkova, & Hulme, 2013; Caravolas, Volin, & Hulme, 2005; Frith et al., 1998; Landerl et al., 1997; Mann & Wimmer, 2002; Wimmer & Goswami, 1994; Ziegler, Perry, Ma-Wyatt, et al., 2003). Although such experiments can be very informative in isolating cross-linguistic differences and similarities, as we outlined in the introduction, even a well-matched study is subject to the possibility that uncontrolled or unknown factors drive the cross-linguistic difference, rather than it being attributable to the orthographic characteristic that is assumed to drive the effect.

With this in mind, we chose to test bilingual German/English children to provide a strong test of the PGST. If orthographic depth is the determining factor for a larger body-

N effect in English than German, as reported by Ziegler et al. (2003), we should find the interaction, even when we hold all participant-level characteristics constant by using a within-subject design. If orthographic depth forces children to rely more on bodies in a deep compared to a shallow orthography, the predictions are clear: even in a sample of bilingual children, the orthography should affect the cognitive mechanisms underlying reading. Thus, we should obtain a body-N by language interaction in this within-subject design. If we do not find an interaction between language and body-N, this would suggest that participant-level factors other than orthographic depth drive cross-linguistic differences in reliance on body units which, in turn, would call for future investigation of possible factors.

3.3.1. Method

The participants were 28 students at a German-English bilingual school in Australia. The school followed both the curriculum of the state of New South Wales, and the curriculum of the German state Thuringia. Accordingly, the students had started receiving some reading instruction in English in preschool (approximately at 5 years of age), and German reading instruction from Grade 1 onwards (approximately at the age of 7).

Seventeen children were from Grade 2, six were from Grade 3, and five were from Grade 4. Across the sample, the average age of the children was 8;4 (SD = 0;11, ranging from 7;1 to 10;3). Due to the small number of students from Grades 3 and 4 we collapsed across all year groups.[14]

[14] The number of participants was limited by the number of participating students at the school. We could not obtain more data, since recruiting English-German bilingual children outside of this school would have forced us to compromise on the homogeneity of the sample.

Most children were tested in both English and German, but some students had recently moved to Australia from Germany and did not read English sufficiently well to complete the English tests (4 students from Grade 2 and 1 student from Grade 3) and one had transferred from an English school and could not read enough German to complete the German tests (Grade 4). As LME can handle mixed designs relatively well, the data from the sessions that these children completed were retained in the analyses. All other children were tested in two separate sessions, one for each language. The sessions were at least two days apart (on average, 10.14 days, SD = 3.62). In addition to the body-N items, we assessed the children's German overall reading ability with the SLRT-II 1-minute sight word reading test, and for English we used the sight word reading test from the TOWRE (Torgesen, Wagner, & Rashotte, 1999), which has an almost identical format to the SLRT-II. The outcomes of the sight word reading tests are shown in Table 4. The item presentation for the experimental manipulation was identical to Experiment 1.

Table 4

Sight word reading ability of the children in Experiment 2; mean (SD). For German, SLRT-II, and for English TOWRE.

Language	Raw score	Percentile
German	48.11 (21.15)	42.07 (26.38)
English	62.26 (8.13)	61.51 (22.25)

3.3.2. Results

As for Experiment 1, the items were scored offline with the program CheckVocal (Protopapas, 2007). Again, we used a strict marking criterion where a pronunciation for a nonword was only considered correct when it rhymed with its base word. We removed all non-responses (1.21% of all data). The accuracy and latency summaries are shown in Table 5.

Table 5.

Body N effects for words and nonwords in English and German; mean (SD).

	Reaction time (ms)				Accuracy (%)			
	Words		Nonwords		Words		Nonwords	
	High BN	Low BN	High BN	Low BN	High BN	Low BN	High BN	Low BN
German	773.2 (106.5)	796.6 (99.4)	881.8 (78.6)	885.5 (120.2)	89.76 (7.61)	83.17 (18.21)	82.32 (12.77)	75.73 (17.28)
English	718.0 (75.5)	735.7 (85.8)	882.8 (93.4)	913.4 (119.4)	95.41 (8.46)	90.81 (7.26)	85.55 (12.06)	7638 (16.94)

We performed LME analyses with language (German, English), body-N (high, low) and lexicality (words, nonwords) as independent variables. In accuracy, there was a significant effect of lexicality, $z = 5.46$, $p < 0.0001$, with words read more accurately than nonwords, and a main effect of body-N, $z = 3.01$, $p < 0.005$, with a higher accuracy for high than low body-N items. The interaction between language and lexicality approached significance, $z = 1.92$, $p = 0.05$, reflecting a slightly larger lexicality effect in English (12.13%) than German (7.44%). All other $p > 0.2$. Importantly, the interaction between language and body-N did not approach significance, $z < 1$.

For the latency analyses, we used inverse RTs. As a Q-Q plot showed an approximately normal distribution, we did not exclude any outliers. The latency analyses showed a significant main effect of lexicality, $t = 9.79$, $p < 0.0001$, as words were read faster than nonwords. There was a main effect of language, $t = 2.25$, $p < 0.05$, as English items were read faster than German items, and an effect of body-N, $t = 2.18$, $p < 0.05$, where high body-N items were read faster than low body-N items. All other $t < 1.1$.

As the interaction between language and body-N is theoretically important (Ziegler & Goswami, 2005), we followed up on the non-significant results in RT and accuracy with a BF analysis. Both for accuracy and RTs, we contrasted a model

containing the main effects but no body-N by language interaction, to identical models which included this interaction. In accuracy, the BF value was 13.21 (±8.02%), thus providing strong support for the H_0 model against the H_1 model, suggesting that there is no difference in the size of the body-N effect in English and German. For RTs, the corresponding value was 10.38 (±4.38%), thus providing almost strong evidence for the H_0, that the size of the body-N effect does not differ in RTs.

Inverse RT analyses can sometimes mask interactions. Therefore, we repeated the RT LME analyses, while using the z-scores for the RTs, calculated using the by-subject means and SDs. Again, the body-N by language interaction did not approach significance, $t = 0.74$. Using zRT scores, the BF value for the null hypothesis was 13.77 (±4.89%), again showing strong evidence for no interaction.

We conducted an additional follow-up test: although our choice of bilingual participants was deliberate, it could be argued that this sample is not representative of typical readers in either language. To ensure that bilingual children do not behave in unexpected ways, we compared the grade-two children reading the German items to the grade-two German monolingual children from Experiment 1. The LMEs showed a marginally significant difference in overall accuracy, $z = 1.86$, $p = 0.06$ (in latencies, $p > 0.9$), suggesting that monolingual children were overall more accurate than bilingual children. Importantly, the status as a bilingual or monolingual child did not interact with body-N, both for accuracy and RT $p > 0.1$. The BF provided strong evidence for a model which excluded the bilingual-monolingual by body-N interaction, for accuracy, BF = 20.78 (±7.85%), and for RT, BF = 13.43 (±3.75%).

3.3.3. Discussion

In Experiment 2, we aimed to expand on the findings of Ziegler et al. (2003) by assessing the body-N effect across two languages varying in orthographic depth, namely

German and English, in a within-subject design. Furthermore, we matched our items across languages on body-N and orthographic N. This puts a key assumption underlying the PGST under a strong test: if orthographic depth is the driving force behind stronger reliance on body units in deep orthographies, then keeping everything else constant, the body-N effect should be stronger for English items than for German items. We find evidence against such an interaction. In a matched item set, and using a within-subject design, the size of the body-N effect is approximately equal across the two languages in the current experiment.

There are several reasons why our results may be different from those of Ziegler et al. (2003). Specifically, we controlled for several additional item- and subject-level factors. Our items were matched across languages on orthographic density factors, such that the strength of the body-N manipulation did not differ for English and German. We also used a within-subject design. This means that we controlled for factors that are not related to orthographic depth, that may vary across samples, such as reading instructions, parent involvement, and overall cognitive processing efficiency.

As the rationale of using bilingual children to test for cross-linguistic differences in novel to this area of study, several words of caution are due here. Firstly, although we compared the bilingual children reading German to the German monolingual readers of the same age and found no difference in the overall pattern, it is possible that knowledge of a different orthography shapes the way in which words are processed. We consider this unlikely: a recent study with adults has found that the sublexical processes were specific to the orthography rather than knowledge of an additional language or the status of the language as a native or non-native language (Schmalz et al., 2014, i.e., Paper 4).

Secondly, a within-subject design eliminates possible confounds, but it does not guarantee that the children will be matched on their English and German reading ability.

In fact, the reading ability of the children was higher in English compared to German, as shown both by their raw and their standardised scores on the speeded word reading tasks. This is likely due to the earlier onset of reading instruction in English than in German, as well as the fact that the children lived in an English-speaking country. This is in contrast to previous studies which compare German to English children, as it is well established that reading acquisition is slower in English than German (Seymour et al., 2003). Although this might mean that the overall faster reaction times in English masked a relatively stronger body-N effect in English than in German, this is an unlikely explanation for our null-result when it comes to the language by body-N interaction. We performed an additional analysis, where we used the RT *z*-scores, calculated using the means and SDs of each participant, and replicated the non-interaction both in the LME and the BF analysis.

3.4. General discussion

In the current study, we aimed to address several theoretically important questions relating to the use of body units by developing readers. In Experiment 1, we found stable reliance on body units in German-speaking children for nonwords but not words. From this, we conclude that bodies are processed as sublexical units. The finding that even the youngest group of participants showed this effect further suggests that even in a shallow orthography like German, children learn to rely on these units without explicit instruction. The presence of a body-N effect in German children, and the differential results for words and nonwords, gives us important insights into how bodies are processed, and why they emerge as salient units during reading - as we will discuss in more detail below.

In Experiment 2, we tested bilingual English/German children on the body-N effect in both languages. These children showed a body-N effect both for words and for nonwords. A possible explanation for the lack of a lexicality interaction is that their

bilingual background may result in overall smaller vocabulary knowledge. Thus, some words were processed by the bilinguals as nonwords, allowing the body-N effect to emerge. As we did not test the children on any vocabulary measures, there may be alternative explanations for this finding, however.

The critical finding from this experiment was the lack of interaction between the body-N effect and language, suggesting that the reliability of the print-speech correspondences of an orthography does not affect reliance on body units. This challenges the results of a key study underlying the PGST (Ziegler & Goswami, 2005; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003).

3.4.1. Theoretical implications of body-N effects

As outlined in the introduction, it is to a large extent unclear how bodies are processed by the cognitive system. This has important theoretical implications. Numerous studies have shown reliance on body units using tasks such as nonword reading paradigms (will participants pronounce "dalk" as /do:k/ or /dælk/? E.g., Brown & Deavers, 1999; Glushko, 1979; Schmalz et al., 2014) or body-existence manipulations (is "dake" easier to read than "daik"? E.g., Goswami et al., 1998; Goswami et al., 1997; Goswami et al., 2003; Treiman et al., 1990). The conclusions and implications of these studies depend on whether bodies are considered to be sublexical units (Patterson & Morton, 1985), or whether reliance on larger sublexical units is proposed to reflect some aspect of the orthographic lexicon (Forster & Taft, 1994; Goswami, 1991).

From a theoretical perspective, it is important to establish what reliance on bodies means, as this has strong implications for all models of reading. There is a general consensus amongst computational models that a non-lexical procedure is also required to compute the pronunciation of unfamiliar words (M. Coltheart, Curtis, Atkins, & Haller, 1993; M. Coltheart et al., 2001; Perry et al., 2007; Perry, Ziegler, & Zorzi, 2010; Plaut,

1999; Plaut et al., 1996). The exact nature of the sublexical process, however, is a source of disagreement amongst computational modellers. Establishing that bodies are processed by the sublexical route therefore provides a marker effect that can be simulated by computational work in the future.

3.4.2. Bodies across age and orthographies

As we did not find any differences across grades in the reliance on different units, our results cannot be used to support either the large-units first (Goswami, 1993, 2002) or the small-units first hypothesis (Hulme et al., 2002; Nation et al., 2001). Given the stable effect for nonwords across all grades of Experiment 1, however, we can conclude that the sublexical skills of German children are already well-refined in the earliest age group (Grade 2) - at least given the relatively frequent spelling patterns that we used.

Our results also suggest stability of the body-N effect across orthographies. Amongst the same children, the body-N effect did not differ regardless of whether they read English or German items. Furthermore, the size of the body-N effect in German did not differ in monolingual compared to bilingual children, thus making it unlikely that knowledge of another orthography significantly affects the processing underlying reading in bilingual children.

The bilingual design aimed to put the key assumption of the PGST - that the choice of units differs as a function of the orthography's depth - under the strongest possible test. Using a within-subject design, we held all participant-level factors constant, and with a well-matched item set we also excluded potential confounds such as orthographic density or the strength of the body-N manipulation across orthographies. The results suggest that orthographic depth has a minimal influence on the choice of units in children - if any at all.

Future research is needed to address the discrepancy of our results and those of Ziegler et al. (2003). As there were several methodological differences between the two studies, this offers several plausible explanations. First, the language by body-N interaction found by Ziegler et al. (2003) might be due to a cross-linguistic confound with reading instruction. It is likely that the Australian children in their sample were receiving whole word reading instruction, while the German children were receiving phonics reading instruction. Evidence suggests that this is an important factor in determining reliance on graphemes as opposed to bodies in nonword reading tasks (Thompson et al., 2009), therefore cultural differences in the type of reading instructions could be an alternative explanation for the cross-linguistic differences reported by Ziegler et al. (2003).

The second possible explanation relates to our use of a well-matched item set. A cognates design provides control over semantic variables and number of letters and phonemes, but cannot control for other lexical and sublexical variables that differ systematically across orthographies. Therefore, a set of cognates can - and in the case of Ziegler et al. (2001 & 2003) does - differ in terms of orthographic N and body-N. Of particular concern to their conclusion is the stronger body-N manipulation in English compared to German, given that their body-N manipulation was stronger for the English than for the German item set.

3.4.3. Conclusion

In the current study, we report two experiments investigating how bodies are processed, and how reliance on body units differs throughout the process of reading acquisition, and across orthographies. We show that bodies are processed as sublexical units. This has implications for interpreting previous experiments on the use of larger-than-grapheme units. Furthermore, it provides a benchmark for models of reading: the

sublexical route of a computational model needs to be able to show sensitivity to body units. In Experiment 1, we found a stable body-N effect for nonwords across age, suggesting that the sublexical systems of the children are already well-refined. In Experiment 2, we also showed that the nature of the orthography does not affect the reliance on body units as there is no cross-linguistic difference in the size of the body-N effect, thereby challenging the key assumption of the PGST (Ziegler & Goswami, 2005).

Appendix: Items used in Experiments 1 and 2

English

Words		**Nonwords**	
High body-N	Low Body-N	High body-N	Low Body-N
blend	beef	bly	barsh
blink	bird	brug	berge
block	blond	chy	berm
bride	burn	clop	chyle
chest	buzz	clust	clize
clock	cloud	dree	deef
draw	count	drell	dilm
drink	dense	drite	dize
drop	desk	fice	dresh
drum	dish	fride	ferb
dry	egg	frink	fich
free	film	glain	floud
glide	firm	glump	frep
ice	flesh	kleed	frict
ink	flu	klin	gresh
joke	fresh	kump	gule
jump	fruit	loke	gurt
loop	fuss	plide	guss
nice	germ	plim	hish
plug	guy	pling	krilk
plum	harsh	plit	kuy
plump	herb	plock	lerm
pride	hurt	ploke	lodd
quit	juice	preep	stond
skin	loud	prell	lount
sleep	merge	quink	luice
slide	milk	shale	mense
slug	mix	shide	mesk
smell	odd	shoop	murse
smoke	porch	smee	pext
speed	prize	smest	pilk
spring	purse	splaw	pish
spy	quiz	stend	pliz
stain	rich	stide	plu
stop	rule	stug	pluit
straw	shelf	stum	rix
stuff	silk	swuck	roft
stun	size	swuff	roud
swim	soft	traw	sird
tree	step	trest	stelf

troop	strict	trock	suzz
truck	style	trop	tirm
trust	term	whun	wurn
whale	text	wrum	yorch
white	wish	zoop	zegg

German

Words		**Nonwords**	
High body-N	Low Body-N	High body-N	Low Body-N
acht	Aal	bast	beld
Bank	Amt	grein	damt
blank	Angst	gund	dels
blau	bald	jand	dinn
Brand	Boot	jaus	faub
Dank	Busch	jur	femd
Dreck	eins	klank	fold
ein	Feind	klund	gald
Fang	Feld	klur	gaub
fast	Fels	krast	goo
Fleck	Film	krau	goos
Flur	fünf	kreck	jald
Fracht	Gold	kreil	jenf
Frau	Graf	lank	kangst
Fang	halb	mang	keiz
Fast	Heft	mank	kust
grau	Held	nast	laat
Frund	Helm	pang	lönch
Fund	Hemd	pau	lünf
Facht	Holz	placht	lusch
klein	Kinn	pland	malb
krank	Kohl	plaus	melz
kraus	Kreuz	plein	mohl
Fur	Laub	plur	naal
lang	links	prank	nech
macht	Lust	prau	nehr
Maus	Mehl	preck	paf
Pfeil	mehr	preil	pah
Pfund	Mönch	quang	pehl
Pracht	Moos	schur	pelm

Rank	nah	schweck	pies
Sau	Netz	spand	poot
Schnur	Pech	krand	reft
Speck	Pelz	spau	retz
Spur	Raub	splast	reuz
Stand	Reiz	splur	rinks
Stau	Saat	sprau	seind
steil	Samt	spund	seng
Stein	Senf	stacht	silm
Strand	Spieß	gracht	sohl
stur	streng	tacht	tehn
Tracht	stück	treck	teins
und	Wald	trur	teld
Wand	zehn	wank	truck
Zweck	Zoo	zacht	wolz

Paper 4: Quantifying the degree of reliance on different sublexical correspondences in German and English

This paper is now published (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. doi: 10.1080/20445911.2014.968161).

Abstract

The type of the sublexical correspondences employed during nonword reading has been a matter of considerable debate in the past decades of reading research. Nonwords may be read either via small units (graphemes), or large units (orthographic bodies). In addition, grapheme-to-phoneme correspondences may involve context-sensitive correspondences, such as pronouncing an *a* as "/ɔ/" when preceded by a *w*. Here, we use an optimisation procedure to explore the reliance on these three types of correspondences in nonword reading. In Experiment 1, we use vowel length in German to show that all three sublexical correspondences are necessary and sufficient to predict the participants' responses. We then quantify the degree to which each correspondence is used. In Experiment 2, we present a similar analysis in English, which is a more complex orthographic system.

4.1. Quantifying the reliance on different sublexical correspondences in German and English

How print is converted to speech is an important question, both from a theoretical and practical perspective. Sublexical translation processes have a central role in all current models of reading aloud (M. Coltheart et al., 2001; Perry et al., 2007; Perry, Ziegler, & Zorzi, 2010; Plaut et al., 1996; Seidenberg & McClelland, 1989). The exact nature of this sound-to-speech conversion procedure, however, has been under considerable debate since the 1970s. In particular, the debate revolves around the question of whether this conversion relies predominantly on small units, such as graphemes, or larger units, such as orthographic bodies (e.g., *-ord*) (Andrews, 1982; Glushko, 1979; Jared, 2002)[15]. To a lesser extent, the literature has also drawn a distinction between context-sensitive and context-insensitive grapheme-to-phoneme correspondences (GPCs) and addressed the possibility that rather than relying purely on single-grapheme correspondences, in some cases the preceding or succeeding letters may provide a cue to the reader about the correct pronunciation of a grapheme (Perry, Ziegler, Braun, et al., 2010; Treiman et al., 2003; Treiman, Kessler, Zevin, Bick, & Davis, 2006).

Thus, the literature reports three different types of correspondences that may be involved in sublexical decoding: context-insensitive GPCs, context-sensitive GPCs, and body-rime correspondences. Here, we propose a mathematical model based on an optimisation procedure that will allow us to fit the degree of reliance on each of the three types of correspondences. We begin with two experiments in German, where the language structure allows us to assess the independent contribution of each of the three types of correspondences. In two further experiments, we apply the same methodology to

[15] It is not always true that graphemes are smaller (i.e., contain fewer letters than) bodies, e.g., the grapheme *igh* is larger than the body of the word *cat* (*-at*). For the sake of clarity, we follow the terminology of Ziegler and Goswami (2005) and refer to graphemes as small units, and bodies as large units.

the English grapheme *a*, which allows us to disentangle the reliance on context-sensitive GPCs compared to context-insensitive GPCs.

GPCs describe the relationship between graphemes and phonemes. The phoneme is the basic unit in spoken language, and a grapheme is the letter or letter cluster that corresponds to a single phoneme. The definitions of GPCs are straightforward in some cases; for example, the grapheme *b* always maps onto the phoneme /b/. This is an example of a context-insensitive GPC: regardless of the letters that precede or succeed the grapheme, its assigned phoneme does not change. However, this gets more complicated when we consider the GPC for a grapheme such as *a*. In English, context-insensitive correspondences would dictate that *a* should be pronounced as in "cat". Using this correspondence, words like *was* and *false* would be considered irregular, meaning that the correct pronunciation is inconsistent with the GPC. Yet, upon closer inspection, the pronunciations of *was* and *false* are entirely predictable when the context of the grapheme *a* is taken into account: in *was*, the *a* is preceded by a *w*, which in most cases changes the pronunciation to /ɔ/, as in "wad" and "swan".[16] This context-sensitive correspondence can be written as *[w]a* → /ɔ/, where an *a* is pronounced as in "bald" when preceded by a consonant, and followed by an *l* and another consonant (hereafter: a[l]-correspondence). It is worth noting that these context-sensitive correspondences are still GPCs, as they relate a single grapheme (in this case, *a*) to the pronunciation of a single phoneme. Thus, GPCs can be subdivided into context-sensitive GPCs (*[w]a* → /ɔ/) and context-insensitive GPCs (*a* → /æ/).

The concept of GPCs is important for the classical computational model of the dual-route framework, the DRC (M. Coltheart et al., 2001). This model has a sublexical

[16] There are some differences associated with dialects. Here, we use the pronunciations given by the DRC's vocabulary and the Macquarie Essential Dictionary (5th Edition) as representative of Australian English, and the IPA as illustrated by Cox and Palethorpe (2007)

route which converts print to speech via a set of GPCs that are explicitly specified. The sublexical route contains some context-sensitive correspondences (N = 28 - though the exact numbers vary according to the version of the DRC), but operates mostly on single-letter (e.g., *b* → /b/; N = 40) and multi-letter (e.g., *th* → /θ/; N = 165) context-insensitive GPCs.

There is also experimental evidence that stresses the importance of context-sensitive correspondences. One study reported the case of a patient with acquired surface dyslexia (Patterson & Behrmann, 1997): since this patient could not correctly read irregular words like *colonel* and *yacht*, it was thought that her lexical system was heavily damaged. However, not all irregular words were a problem: she was unimpaired with words that could be resolved by the context-sensitive *[w]a* → /ɔ/ correspondence, such as "wad" or "swan". This demonstrates the presence of such a context-sensitive correspondence in the sublexical system. Furthermore, studies of nonword reading have shown that there is psychological reality to context-sensitive correspondences (Treiman et al., 2003; Treiman et al., 2006): both adults and children tend to pronounce nonwords such as TWAMP with the vowel as in "swan", whereas control items such as GLAMP are pronounced via the context-insensitive GPC, *a* → /æ/. This further suggests that the context-insensitive correspondence *a* → /æ/ does not fully reflect the strategies used during nonword reading.

In addition to context-insensitive and context-sensitive GPCs, readers have been shown to rely on body-rime correspondences. Body-rime correspondences are the sublexical links between bodies and rimes, where bodies are defined as the vowel and optional final consonant(s) of a monosyllabic word (e.g., *-ark* in the word *bark*). The rime is the phonological equivalent to the orthographic body. A linguistic analysis has shown

that bodies are a reliable predictor of vowel pronunciation in English (Treiman et al., 1995).

Full reviews of the psychological reality of body-rime correspondences can be found elsewhere (Goswami & Bryant, 1990; Ziegler & Goswami, 2005). Most relevant in the current context are nonword reading studies addressing this issue, because these allow for a systematic exploration of the non-lexical correspondences that participants rely on when lexical information is not available. In English, nonwords can be created that would yield different responses depending on whether GPCs or body-rime correspondences are used. This is done by manipulating the regularity and consistency of the base word. A base word that conforms to the context-insensitive GPCs is said to be regular, while words that violate the correspondences are considered to be irregular. The concept of regularity only matters if reading occurs at least in part via GPCs. If nonlexical reading occurs only via body-rime correspondences, then the reliability (or lack thereof) of the GPC information should not influence reading at all; rather, only inconsistency of body-rime correspondences should affect reading latencies and accuracy (e.g., the two ways of pronouncing -ave in "have" and "save"; see Ziegler et al., 1997).

Nonword reading studies that aim to estimate the reliance on GPCs versus body-rime correspondences can use the regularity and consistency of a base word to generate nonwords that predict different responses depending on the types of correspondences that are used by the participant. Such studies are important, because nonword reading data can shed light on processing underlying sublexical information, while minimising confounds from lexical processing. Understanding this process has strong theoretical implications, because sublexical print-to-speech conversion mechanisms play an important role in all prominent models of reading.

In order to disentangle the different sublexical processes that take place during reading, the first step is to create nonwords for which different types of correspondences make different predictions. For example, from a regular and consistent word such as *fact*, the onset can be changed to create a nonword, for example, RACT. In this case, both large and small correspondences make the same predictions for pronouncing this nonword. However, if we take an irregular, but consistent word, such as *talk*, and change the onset to create the nonword RALK, we can use the readers' pronunciations of this nonword to determine whether they relied on context-insensitive GPCs (in which case the item would be pronounced to rhyme with "talc") or body-rime correspondences (where it would rhyme with "talk"). Such studies have shown that GPCs cannot fully account for the types of pronunciations that participants give to such nonwords, but neither do body-rime correspondences (Andrews & Scarratt, 1998; Brown & Deavers, 1999; Perry, Ziegler, Braun, et al., 2010; Pritchard et al., 2012).

Thus, there is evidence for reliance on the three different types of print-to-speech correspondences, but there are still questions that remain to be answered. First, previous studies do not distinguish between the reliance on context-sensitive GPCs and body-rime correspondences. For example, if a participant pronounces the nonword PALSE to rhyme with "false", it may be that a context-sensitive correspondence, *a[l]* → /oː/ has been used to derive the pronunciation, rather than the BRC that *-alse* → /oːls/. As will be discussed later, this is a problem in the English language, as body-rime correspondences and context-sensitive correspondences are confounded.

Second, even though such studies can establish the psychological reality of certain types of correspondences, examining between-item differences does not allow estimation of the *relative degree* to which each type of correspondence plays a role. As previous research has demonstrated the psychological reality of context-insensitive GPCs, context-

sensitive GPCs, and body-rime correspondences, it is likely that all three correspondence types help the sublexical route to determine the pronunciation of a nonword. How such a conflict between different types of correspondences may be resolved by the cognitive system is addressed in detail in the General Discussion. The possibility of parallel activation of several sublexical correspondences raises the question of whether it is possible to quantify the degree to which each plays a role in determining the pronunciation of a novel word, which is a natural next step after demonstrating a sublexical correspondence's psychological reality. As discussed below, more sophisticated analyses are needed to estimate the relative importance of each type of correspondence.

In addition to establishing the psychological reality of different types of sublexical correspondences, a considerable body of research has explored cross-linguistic differences in the reliance on GPCs versus body-rime correspondences (Goswami et al., 1998; Goswami et al., 1997; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). The psycholinguistic grain-size theory, a cross-linguistic theory of reading development and skilled reading, proposes that the degree of reliance on sublexical correspondences of different types varies across languages (Ziegler & Goswami, 2005). In particular, the reliance on body-rime correspondences has been reported to be stronger in English than German (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). This is argued to be because in English, large units (i.e., bodies) are a better predictor of the pronunciation of a word than GPCs (Treiman et al., 1995): for a word like *calm*, the pronunciation is inconsistent with the GPCs ("/kælm/") but can be derived from its body neighbours (palm, balm, etc.). In German, on the other hand, the GPCs are highly reliable, meaning that there are few exceptions to the correspondences (Ziegler et al., 2000), therefore smaller units are the preferred grain-size of German readers. In other words, there is a

theoretical framework which predicts differences in the reliance on the units across languages. Therefore, it is desirable to develop a mathematical model quantifying the degree of reliance in different languages.

In summary, previous literature has shown reliance on three different types of correspondences in English: context-insensitive GPCs, context-sensitive GPCs, and BRCs. The psycholinguistic grain-size theory proposes that the reliance on the different types of correspondences differs across languages (Ziegler & Goswami, 2005). In the present experiments, we introduce a new method of quantifying the reliance on each type of correspondence. In the first two experiments (1A and 1B), we used German nonwords to assess the degree of reliance on each type of correspondence. In Experiment 2A and 2B we extend the procedure to a more complex orthographic system, namely English.

4.2. Experiment 1A

The German language allows us to neatly assess the independent contributions of context-insensitive GPCs, context-sensitive GPCs, and body-rime correspondences in a nonword reading paradigm: It is possible to create a set of items which generate different predictions for vowel pronunciation, depending on which strategy is used.

In German, there is relatively little ambiguity in print-to-sound correspondences, compared to English. What little ambiguity there is stems mostly from vowel pronunciation (Ziegler et al., 2000): Each vowel can be pronounced as either long or short (e.g., *Schal* → "/ʃaːl/" versus *Schall* → "/ʃal/"). In monosyllabic words, vowel length is often signalled by context. Some context-sensitive correspondences allow the reader to unambiguously determine vowel length; for example, any vowel followed by an *h* is pronounced long (*V[h]* → "long vowel"). Other context-sensitive correspondences are less transparent. These correspondences are described by a German implementation of Coltheart et al.'s (2001) DRC (Ziegler et al., 2000). To allow the sublexical route to

determine vowel length, it contains a set of context-sensitive super-rules: any vowel which is followed by only one consonant elicits a long vowel response (e.g., *Wal*), and a vowel which is followed by two or more consonants is pronounced short (e.g., *Wald*). These two rules can be summarised as follows: *V[C]* → "long vowel", and *V[C][C]* → "short vowel".

Although these super-rules capture the overall statistical distribution, there are also some exceptions, or words that would be irregular according to the German DRC (Ziegler et al., 2000). The word *Magd*, for example, is pronounced with a long vowel; conversely the word *Bus* is pronounced with a short vowel. The presence of several bodies which consistently break the super-rules allows us to orthogonally manipulate the number of consonants in the body of a nonword, and the pronunciation of the base-words. Thus, we create a situation where the different types of correspondences (i.e., super-rules and body analogy) make different predictions about the pronunciation of the vowel.

For the present experiment, we can make a set of simple predictions if we assume that readers generally use only one type of correspondence: If only context-insensitive GPCs are used for German nonword reading, we expect that the likelihood of a short vowel pronunciation should be independent of any other orthographic features of the nonword. Such a GPC would predict many more short than long vowels, as the majority of vowels in German have the short pronunciation (Perry, Ziegler, Braun, et al., 2010). If a context-sensitive super-rule is used, vowel length should be solely determined by the number of consonants following the vowel. In this case, even a nonword based on the irregular but consistent word such as *Magd* (e.g., BLAGD) should be pronounced with a short vowel. These irregular-base-word items can distinguish between reliance on super-rules compared to body-rime correspondences: if body-rime correspondences are used,

nonwords based on irregular consistent words should be pronounced to rhyme with their real-word counterparts.

4.2.1. Methods

Participants were 12 German native speakers who were staff or postgraduate students at Macquarie University, or members of the university's German society. As they lived in Australia, they were also fluent in English - a point which we will discuss in a later section. With one exception, all participants had completed secondary education in Germany and 10 had also attended German tertiary education. One participant had moved to Australia at the age of 5, but had attended a German-speaking school for 7 years.

The nonwords that were used for this experiment are listed in Appendix A. There were 30 nonwords in each of three conditions. The nonwords were created by changing the onsets of real words. All base-words were taken from a list of consistent German words (J. Ziegler, personal communication, 2012). The first condition used base-words with *V[C]* bodies which were pronounced with a long vowel (*Jod* → FOD) ; the second condition was based on *V[C][C]* words with a short vowel (*Saft* → BLAFT). The third condition was derived from irregular words, which had either a *V[C]* body but a short vowel (*mit* → GIT) or a *V[C][C]* structure and a long vowel (*Jagd* → BAGD). The three conditions were matched on orthographic N (the number of real words that can be created by substituting one letter): *V[C]* items had an average orthographic N-size of 1.73 (SD = 1.46), *V[C][C]* items had a mean of 2.10 (SD = 1.69), and items with irregular base-words had a mean of 1.83 (SD = 1.90). The mean body-N (number of real words with the same body) for the three conditions is 1.93 (SD = 1.87), 2.40 (SD = 1.98), and 1.37 (SD = 1.00) respectively.

Participants were tested individually in a quiet room. Instructions were given in German by a native speaker. The participants were told that they would be asked to read

nonwords which were created using German orthographic rules. The instructions emphasised that accuracy was more important than speed to discourage quick lexical processing, which might result in lexicalisation errors.

The items were presented using the DMDX software package (Forster & Forster, 2003) in random order. Each trial consisted of a fixation cross, which remained in the centre of the screen for 500 ms, followed by the item, which remained on the screen until the voice-key was triggered. Ten practice nonwords preceded the experiment. As all nouns in German are spelled with capital initial letters, presenting nonwords in all lower-case would provide an indication of word class of a nonword. Previous research has shown that information on the likely word class of a nonword affects its pronunciation (Campbell & Besner, 1981). Therefore, all items were presented in upper case.

4.2.2. Results

Six trials (0.6%) were excluded due to poor sound quality or premature voice-key triggering. The rest of the trials were scored by a German native speaker as pronounced with a long vowel, a short vowel, or incorrectly. For identifying incorrect responses, we used a lenient marking criterion: if a participant's response was consistent with a possible pronunciation of the GPCs, it was marked as correct (e.g., *spic* was marked as correct regardless of whether it was pronounced as "/spik/" or "/ʃpik/") - while, in German, *s* is typically pronounced as "/ʃ/" before *p* or *t*, there are a few instances, such as loanwords, where it is assigned the pronunciation "/s/"). Overall, 1.1% of all responses were classified as incorrect and excluded from subsequent analyses.

Of primary interest were the proportions of long and short vowel responses and how they differed across condition. We split the Irregular-base-word condition by whether the bodies had one (hereafter referred to as *V[C]* Irregular; N = 17) or two (*V[C][C]* Irregular, N = 13) consonants. Note that the two "irregular" conditions did not

differ dramatically on any item characteristics: the mean number of letters was 3.88 (SD = 0.34) and 4.36 (SD = 0.63) for the V[C] and V[C][C] conditions respectively, orthographic N was 1.88 (SD = 2.19) and 1.79 (SD = 1.58) respectively, and body-N was 1.56 (SD = 1.15) and 1.29 (SD = 0.61) respectively. The proportions of short vowel responses for each of the four item types (V[C] Regular, V[C][C] Regular, V[C] Irregular and V[C][C] Irregular) are listed in Table 1, along with the predictions according to each of the three types of correspondences.

Table 1.

Percentage of short vowel responses for each condition in Experiment 1 and the average predictions from each of the three types of correspondences.

	V[C] Regular	V[C][C] Regular	V[C] Irregular	V[C][C] Irregular
Example	*"bral"*	*"brald"*	*"brus"*	*"bragd"*
% Short 1A	47.25	83.69	84.63	61.04
% Short 1B	37.28	86.79	72.95	62.84
Correspondence predictions				
P(Short|GPC)	70.21	79.53	90.59	78.77
P(Short|CSC)	26.20	92.57	62.82	91.38
P(Short|BRC)	2.76	100.00	100.00	0.00
Model predictions				
% Short 1A	44.68	87.04	87.83	60.53
% Short 1B	36.58	89.78	83.67	61.70

GPC, context-insensitive GPC; CSC, context-sensitive GPC; BRC, body-rime correspondence.

In order to make the predictions more specific, we can use a corpus analysis to determine the percentage of times a vowel is pronounced as long or short under certain circumstances. For example, overall, 78.02% of all monosyllabic words are pronounced

with a short vowel (Perry, Ziegler, Braun, et al., 2010) therefore if German readers rely on context-insensitive GPCs, we expect them to give around the same percentage of short vowel responses. Among words with a single-consonant coda, 24.53% are pronounced with a short vowel, so we expect about the same percentage of short vowel responses to V[C] nonwords, if only super-rules are used to determine vowel length. In Table 1, we present the predicted vowel lengths for each of the four conditions and by each of the three types of correspondences. For the context-insensitive GPCs and super-rules, these are calculated from the analyses presented in Perry et al. (2010). The predictions of the body-rime correspondences depend on the consistency ratio of the body. In the current study, we used only consistent items, where the body has only one pronunciation in real words. This means that if participants rely solely on body-rime correspondences, 100% of the pronunciations should be consistent with the base word vowel length.

The obtained percentages of long and short vowels (Table 1) are not consistent with the predictions of any one strategy we described above: vowel length responses are neither predominantly short in all four conditions, nor completely dependent on the number of consonants following the vowel, nor the vowel length of the base word. This is a clear indication that German readers rely on more than one type of correspondence for reading nonwords. Moreover, a closer look at Table 1 shows that no combination of two types of correspondences can account for the results, either: If context-insensitive GPCs and context-sensitive correspondences were the sole determiners of vowel length, we would not expect to find different proportions for the V[C] Regular and V[C] Irregular items - but we do. If context-insensitive GPCs and body-rime correspondences were the only predictors of vowel length, we would find no difference between the V[C] Regular and the V[C][C] Regular items - and we do. If only context-sensitive correspondences

and body-rime correspondences were used, we should observe less than 25% short vowel responses - which is not supported by the data.

4.2.3. Modelling vowel pronunciations

It is not possible for a single or even a pair of types of correspondences to adequately fit the empirical data. It may be, however, that some combination of all three types of correspondences provides a good fit. Here we introduce a mathematical modelling approach that allows us to uncover more complex relationships between the types of correspondences. The goal is to weight the three strategies[17] in a way that optimally fits the empirical data. More formally, we are seeking a set of β weights that best satisfy the following mathematical model (one pair of equations for each item):

$$P_i(Short) = \beta_{gpc} \times GPC_{short,i} + \beta_{csc} \times CSC_{short,i} + \beta_{brc} \times BRC_{short,i} \quad (1)$$

$$P_i(Long) = \beta_{gpc} \times GPC_{long,i} + \beta_{csc} \times CSC_{long,i} + \beta_{brc} \times BRC_{long,i}$$

where $GPC_{length,i}$ is the probability of item *i* being pronounced with a vowel of the corresponding length according to the corpus analysis when using only context-insensitive (single-letter) GPCs as a predictor, $CSC_{length,i}$ is the probability according to context-sensitive super-rules, and $BRC_{length,i}$ is the probability according to the BRCs. Table 1 provides the average predictions for each condition, but the predictions from each correspondence were calculated separately for each item in the experiments. $P_i(length)$ is the empirically observed proportion of the vowel length in Experiment 1A.[18]

[17] Though we refer to the reliance on different types of correspondences as a "strategy", we do not mean to imply that readers *consciously* choose the type of correspondence that maximises the chance of correctly reading an unfamiliar word.

[18] In standard linear regression, only one of these two formulae would be required, since they are entirely dependent (i.e., $P_i(long)=1-P_i(short)$, etc...). In traditional regression, the only difference between the first and second equations would be the location of the estimated intercept and the sign of the slope. However, by removing the intercept term, our modelling strategy undermines this interdependence. Since the intercept is not free to vary (it is forced to be 0) the parameter estimates for P(short) would not match those for P(long). As a result, we must simultaneously fit both vowel pronunciations. While it is

At a first glance, this would appear to be a simple regression problem (with no intercept term). Linear regression would optimally select β values that minimised the prediction error for (1) (indexed by the residual sum of squares). However, there are several reasons why this should not be thought of as regression. First, since the β values are thought of as the degree to which a strategy applies in reading the items in Experiment 1, negative values would be uninterpretable. This means that all of our β parameters must exceed 0. This constraint can not be guaranteed by standard linear regression using ordinary least squares (Monfort, 1995).

Even with only positive βs, there are two ways to interpret the weights. One could think of them as the contribution of each strategy to some sort of blending process that ultimately chooses the vowel pronunciation. In which case, we can simply fit the model in (1) above with the constraint that $\beta_i, \forall i$. Alternately, one can think of the weights as the probabilities of adopting the vowel prediction from a given strategy. We prefer the latter interpretation (and discuss some evidence for it later), but it requires two further constraints: the β weights must fall below 1, and, since we assume the three strategies (GPCs, super-rules, and body-rime correspondences) are exhaustive, the three βs must sum to 1. The model can be formalised as:

$$P_i(Short) = \beta_{gpc} \times GPC_{short,i} + \beta_{csc} \times CSC_{short,i} + \beta_{brc} \times BRC_{short,i}$$

$$P_i(Long) = \beta_{gpc} \times GPC_{long,i} + \beta_{csc} \times CSC_{long,i} + \beta_{brc} \times BRC_{long,i} \quad (2)$$

$$where\ \beta_j \in [0,1]\ and\ \Sigma\, \beta_j = 1, \forall j \in \{gpc, csc, brc\}$$

that is, we are seeking a set of probabilistic weights on the three strategies that minimises the prediction error of the model. The challenge here is to both efficiently search the

---

useful to use the language of regression to describe some of the procedures, it is very important to remember that the βs here do not represent regression slopes, but weights. Also, if this were a regression problem, it would be more properly treated as a *logistic* regression problem. However, this would be incompatible with our interpretation of the weights as "the probability that a certain strategy is adopted."

available parameter space, and satisfy the $\Sigma\, \beta_j = 1$ constraint. The first problem is a well-studied one in computer science and solutions are available that solve it. The second problem is largely solved by introducing an additional equation that can only be satisfied if $\Sigma\, \beta_j = 1$, and giving that equation a strong influence on the final parameter set. The interested reader can find a fuller discussion of the implementation details in Appendix B.

4.2.3.1. Optimal weights in Experiment 1A

In Experiment 1A, we collected the proportion of short and long vowel responses to 90 items, and for each item we have the predicted probability of a short or long vowel pronunciation according to each of the three strategies. The strategy predictions were obtained from the corpus analysis undertaken by Perry et al. (2010).

Using the technique described above, the native German readers in Experiment 1A appear to be relying most on GPCs ($\hat{\beta}_{gpc}$ = 0.56), and to a lesser extent on super-rules ($\hat{\beta}_{csc}$ = 0.19), and body-rime correspondences ($\hat{\beta}_{brc}$ = 0.26). See Table 2 for a summary of the modelling results across all of the present experiments.

Table 2.

Weightings for the three types of correspondences in Experiments 1A, 1B, 2A, and 2B.

Correspondence type	1A (German bilingual)	1B (German monolingual)	2A (English monolingual)	2B (English bilingual)
GPC	0.56	0.38	0.05	0.03
CSC	0.19	0.35	0.69	0.61
BRC	0.26	0.27	0.26	0.36

The above analysis contains a theoretically supported but strong assumption that readers use *only* the three strategies described in the introduction when reading nonwords. It is possible that other sources of information are used by German native speakers to determine vowel length. We can provide a simple test of this possibility by relaxing some

of the constraints on the model, and observing how critical those constraints were to the optimisation results. To do this we removed the $\Sigma\,\beta_j = 1$ constraint, and allowed the βs to take on any positive weights in the fitting process. That is, we fit the following alternative model (some subscripts indicating length and item have been omitted for simplicity):

$$P(length) = \beta_{gpc} \times GPC_{length} + \beta_{csc} \times CSC_{length} + \beta_{brc} \times BRC_{length} \qquad (3)$$

$$where\ \beta_i > 0, \forall i$$

If readers are adopting other strategies that are not well described by the GPC, super-rules and BRC strategies, the incomplete nature of the model should be reflected in these alternate weights. The weights that optimise (2) were $\hat{\beta}_{gpc}$ = 0.58, $\hat{\beta}_{csc}$ = 0.14, and $\hat{\beta}_{brc}$ = 0.24. These values sum to 0.96, suggesting that there is little need for a fourth strategy to describe the data. This does not conclusively rule out a role for any other strategies, but provides some evidence that the three strategies already tested are sufficient. That said, there is one additional strategy that could be playing a role: anti-body correspondences, or the probability of a vowel being pronounced as long or short based on the onset of the word. In this corpus of nonwords, the predictions from ABC and context-insensitive GPCs are highly correlated, so it is difficult to disentangle the two strategies entirely, but it may be that anti-body rime correspondences are more important than context-insensitive GPCs and thus are a better predictor. To test whether or not anti-body correspondences were important for determining vowel pronunciations, we added a component to model (2):

$$P(length) = \beta_{gpc} \times GPC_{length} + \beta_{csc} \times CSC_{length} + \beta_{brc} \times BRC_{length} + \beta_{abc} \times$$

$$ABC_{length}, where\ \beta_j \in [0,1]\ and\ \Sigma\,\beta_j = 1 \qquad (4)$$

where the addition of *ABC* represents the predictions from anti-body correspondences, and $\hat{\beta}_{abc}$ is the associated weight. Fitting (3) produced the same weights that resulted

from (2) where the antibody-rime correspondences were not included. That is, $\hat{\beta}_{abc}$=0, giving little reason to believe that any other strategies are being used in Experiment 1A.

4.2.3.2. Model fits

The optimisation procedure presented here is only useful if it arrives at a model that fits the data better than alternatives. To determine the effectiveness of the model, we calculated the correlation between the model predictions and the observed response patterns. For comparison, we did the same for the GPCs, context-sensitive correspondences (CSC), and BRCs individually. As can be seen in Table 3, the optimisation process outperforms the other three alternatives in all four samples presented here. In experiment 1A, the correlation is .844 while the next best model (based on context-insensitive GPCs) correlates at .714.

Table 3.

Summary of the fits between the models and the observed response predictions. Each value is the correlation between the predictions from the GPC, CSC, BRC, or model and the observed response pattern.

Sample	GPC	CSC	BRC	Optimal (95% CI)
1A (German bilinguals)	0.714	0.681	0.540	0.844 (0.830, 0.847)
1B (German monolinguals)	0.578	0.730	0.659	0.827 (0.812, 0.832)
2A (English monolinguals)	0.522	0.630	0.385	0.729 (0.719, 0.731)
2B (English bilinguals)	0.514	0.573	0.568	0.792 (0.785, 0.793)

4.2.3. Discussion

In Experiment 1A, we successfully used an optimisation procedure to quantify the degree of reliance on three types of sublexical correspondences: context-insensitive GPCs, context-sensitive GPCs, and BRCs. This can be achieved with the German

language, because it is possible to create items where different correspondence types make different predictions about the vowel length pronunciation.

Importantly, we found that all three types of correspondences are both *necessary* and *sufficient* to predict vowel length responses in a sample of German native speakers. Context-insensitive correspondences appear to be the strongest predictor. This is in line with the psycholinguistic grain size theory, which argues that the smallest unit size is favoured by readers of a language with predictable GPCs, such as German (Ziegler & Goswami, 2005).

Experiment 1A has some limitations. It could be argued that the results are unreliable, firstly due to the small sample size and secondly because the participants were bilingual, and very fluent in English. It is unclear how fluency in English may affect the reliance on different types of correspondences in German. Even though we took care to only include German participants who learned to read and write in German from a young age, there is a possibility that their exposure to German reading material has been diminished by residing in an English-speaking country. It is also possible that their knowledge of English would change the preferred unit in their native language: for example, psycholinguistic grain size theory predicts that readers of English rely more heavily on larger grain sizes than readers of German (Ziegler & Goswami, 2005), though it does not make any statements about sublexical processing in bilinguals. We address these concerns in Experiment 1B.

4.3. Experiment 1B

In Experiment 1B we collected data with two different samples of German native speakers who live in Germany and are not exposed to English on an everyday basis. We hereafter refer to them as monolingual Germans, even though they are not strictly monolingual: due to globalisation, it would be difficult if not impossible to find Germans

who have no knowledge of English. Having collected data with two different samples of monolingual Germans allows us to test the reliability of the modelling method described here. If our model arrives at similar weights for two independent samples from the same population, we can be more confident that our modelling procedure is stable and reliable.

4.3.1. Methods

The methods were almost identical to Experiment 1A. One item was replaced (due to a typo, the original item set contained an inconsistent item, BLEN, which was changed to BLEM in Experiment 1B).

The first sample consisted of 10 German native speakers who were staff or students at the Freie Universität in Berlin. All had completed their schooling in Germany. The second sample consisted of 26 undergraduate students at Potsdam University. Again, all were native German speakers and had completed their education in Germany.

4.3.2. Results

The scoring procedure was identical to Experiment 1A. For the Berlin sample, there were two non-responses (0.22%) and 15 errors (1.67%). The Potsdam sample made 2.3% errors. A series of t-tests showed that the percentages of long and short vowel responses did not differ significantly for any of the conditions across the two samples, all $p > 0.4$. Furthermore, fitting each sample separately using the model described in Equation 2 produced very similar weights. For the participants from Berlin, the weights were $\hat{\beta}_{gpc} = 0.40$, $\hat{\beta}_{csc} = 0.33$, and $\hat{\beta}_{brc} = 0.27$. For the participants from Potsdam they were $\hat{\beta}_{gpc} = 0.37$, $\hat{\beta}_{csc} = 0.35$, and $\hat{\beta}_{brc} = 0.28$. This result is comforting, suggesting that the method introduced here is reliable across different samples from similar populations. Since there was little difference between the two samples, we collapsed across them yielding a sample of 36 native German monolinguals. Using this collapsed sample, our

model produces $\hat{\beta}_{gpc}$ = 0.38, $\hat{\beta}_{csc}$ = 0.35, and $\hat{\beta}_{brc}$ = 0.27. As in Experiment 1A, the optimal parameter set outperforms the alternatives in fitting the observed data (Table 3).

4.3.2.1. German/English bilingual versus German monolingual readers

Since Experiments 1A and 1B are based on the same set of items, we have the opportunity to compare how the bilingual readers differed from the monolingual readers. The critical question is whether or not the smaller $\hat{\beta}_{gpc}$ and larger $\hat{\beta}_{csc}$ for monolinguals represents a real difference, or simply random variation. In the usual context of a linear regression model, this would be a simple matter of including the language status of the participants (bilingual vs. monolingual) in the model, and testing for an interaction between language status, and the GPC and/or CSC estimates. However, our modelling strategy violates many of the assumptions that allow for straightforward t-tests of the parameter estimates (given the constraints of our model, the parameter estimates are unlikely to be well-behaved, statistically). Instead we turn to a bootstrapping methodology to allow us to use the data to conduct non-parametric tests of the variability in our estimates.

To establish the reliability of the difference in the $\hat{\beta}_{gpc}$ and $\hat{\beta}_{csc}$ estimates, we repeatedly resampled 90 items (with replacement) from the data set, and estimated the $\hat{\beta}_i$s for both the bilingual and monolingual participants with each sample of items. Of 10,000 such samples, 9,890 (98.9%) produced a larger GPC weight for the bilingual subjects than for the monolingual subjects (95%CI of the difference: 0.019 to 0.327). Similarly, 9,634 (96.2%) samples produced a larger CSC weight for the monolingual participants than for the bilingual participants (95%CI: -0.011 to 0.317). This suggests that the difference in the GPC weights is robust, while the difference in the CSC weights is slightly more tenuous. The difference in the BRC weights was not at all significant: 3,454 (34.5%) of the samples produced larger BRC weights for bilinguals than for

monolinguals (95%CI: -.058 to .089). We also took advantage of these bootstrap samples to estimate the variability in the correlations from the optimal parameters in Table 3.

To summarise the results so far, the reliance on BRCs did not differ between monolingual and bilingual readers, but there was a very stable difference in the reliance on context-insensitive GPCs and a somewhat stable difference in the role of context-sensitive super-rules. Monolinguals relied less on context-insensitive GPCs and somewhat more on super-rules than bilinguals.

4.3.2.2. Individual differences

There is some ambiguity in interpreting the weights: as we collapsed across participants, the weightings do not give us any information about inter-individual participant variability. Theoretically, it is possible that all participants rely on the same strategies to the same extent, or that the weightings are reflective of the percentage of participants who rely on a particular strategy only. To address this, we generated the weighings for each individual participant in Experiments 1A and 1B. These are summarised in Figure 1. This figure shows that there is individual variability, but most participants rely on a combination of the three strategies.

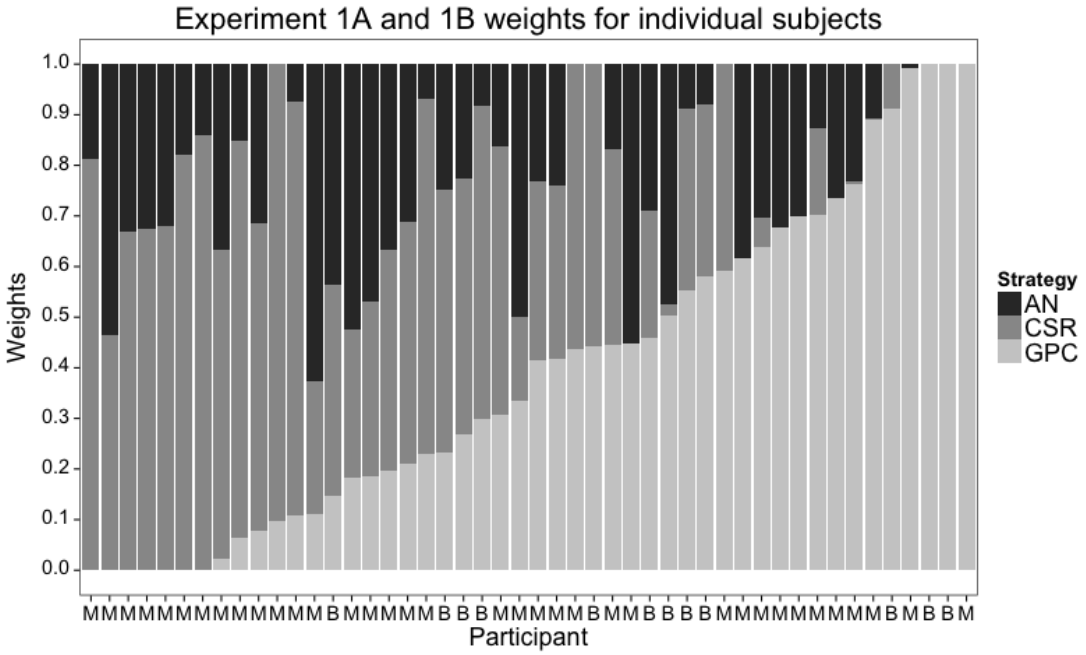

Figure 1. Weightings for each individual participant, sorted by degree of reliance on GPC. GPC - context-insensitive rules, CSR = context-sensitive rules, AN = body-rime correspondences. M = monolingual participant (Experiment 1B), B = bilingual participant (Experiment 1A).

4.3.3. Discussion

As in the previous experiment, we were able to quantify the degree of reliance on each of the three types of correspondences in two samples of monolingual German native speakers. Even though there is individual variation, we found, on average, almost identical reliance on the three strategies in two independent samples of German readers, suggesting that the procedure we introduced is reliable. The overall pattern of results was also broadly consistent with the findings from Experiment 1A, showing that reliance on all three types of correspondences is both necessary and sufficient to explain the vowel length pronunciations in German, and that context-insensitive correspondences are the major predictor of the vowel responses.

While the bilingual and monolingual participants' response patterns were similar, we did find some significant differences in terms of reliance on context-sensitive versus

context-insensitive correspondences: bilingual participants show stronger reliance on context-insensitive correspondences and less reliance on context-sensitive correspondences. Two possible causes of the difference between German/English bilinguals and German monolinguals are the influence of English proficiency on reading in the bilingual sample, or a general difference in German reading proficiency. According to the psycholinguistic grain size theory, if the difference in weights is due to the influence of English (L2) on the choice of correspondences in German (L1), we would expect bilinguals to rely more on larger correspondences (context-sensitive correspondences or body-rime correspondences as opposed to context-insensitive correspondences). Developmental studies have shown that reliance on larger units differs as a function of reading efficiency, as younger children rely to a greater extent on context-sensitive rules (Treiman et al., 2006). In Experiment 1B, we found that bilingual participants rely more on context-insensitive rules, which is more in line with a proficiency explanation - bilinguals may be less proficient in reading German than monolinguals, as they are less exposed to German texts. As a result, they rely to a greater extent on the context-insensitive correspondences. [19]

4.4. Experiment 2A

The majority of prior research on the use of GPCs, context-sensitivity and body-rime correspondences has been conducted in English. In contrast to German, the English

[19] It is noteworthy that Perry et al. (2010) report data with a similar set of nonwords to the current study (though the study was conducted with different aims): the authors manipulated the number of consonants in the coda, but rather than controlling for the consistency of the base-word, their nonwords differed in terms of the existence of the body in real words: the body either occurred in real German words, or it did not. In other words, they did not independently manipulate the predictions of body-rime correspondences and context-sensitive correspondences, and predictions of super-rules and body analogy were heavily correlated, $r(39) = 0.78$, $p < 0.001$, as were the predictions of super-rules and GPCs, $r(39) = 0.51$, $p < 0.001$. This means that the Perry et al. data is unsuitable for our purposes: the analysis would be unreliable, as it is impossible to disentangle reliance on bodies versus super-rules, and super-rules versus GPCs.

letter-to-sound correspondence system is highly complex, as a large set of correspondences on different levels are required to describe the relationship between print and speech (Venezky, 1970). In Experiment 2, we aimed to explore whether it is possible to apply the methodology which we introduced in Experiment 1 to quantify the degree of reliance on the same three strategies in a more complex system.

English, like German, contains some context-sensitive correspondences. However, there are no super-rules, or correspondences which apply to all vowels, as in German. Therefore, we concentrated solely on the grapheme *a*, as its correct pronunciation can often be disambiguated by taking into account its context. By default, *a* is pronounced as in "cat" in Australian English, but there are several context-sensitive and multi-letter GPCs that can modify its pronunciation. The context-sensitive correspondence of interest here is the correspondence that an *a* preceded by a *qu* or *w* is pronounced as "/ɔ/". We chose this correspondence to assess reliance on context-sensitivity for two reasons: Firstly, previous research has shown that there is some psychological reality to this orrespondence (Patterson & Behrmann, 1997; Treiman et al., 2006). Secondly, unlike other context-sensitive GPCs (e.g., *a[l]* → "/oː/") this correspondence is not confounded with body-rime analogy, as the modifier is located in the onset, before the vowel. This is therefore one of the few English context-sensitive correspondences that allows us to independently assess effects of context-sensitivity.

In order to create an item set equivalent to the German nonwords used in Experiment 1, we isolated English bodies with the vowel grapheme *a* which are consistently pronounced irregularly (Ziegler et al., 1997). There are five such bodies: -*alse*, -*att*, -*alk*, -*alt*, and -*ald*. With one exception, they are confounded with the *a[l]* → "/oː/" correspondence: the body -*att* only occurs in the word *watt* and therefore only has the "/ɔ/"-pronunciation. As a result, and in contrast to the German experiment, the degree

of reliance on body-rime correspondences cannot be assessed using this paradigm, because it is almost perfectly confounded with reliance on the *a[l]*-context-sensitive correspondence.

In short, there are three possible pronunciations indicative of reliance on different types of correspondences. If English participants rely on context-insensitive GPCs, we should find that the majority of nonwords are pronounced with the "/æ/"-vowel. If context-sensitive correspondences are used, then in the conditions where a *qu* or *w* precedes the vowel we should find many /ɔ/-responses. If either body-rime correspondences or the *a[l]*-correspondence are used, the conditions with the consistently irregular bodies should be pronounced with an "/oː/".

4.4.1. Methods

The participants were 19 undergraduate students at Macquarie University who were all native speakers of English.

We created four conditions of 18 words each (listed in Appendix A). All were monosyllables containing the single vowel grapheme *a*. The first condition was created by taking consistently regular bodies (Ziegler et al., 1997) and adding an onset which does not change the pronunciation of the vowel (i.e., any onset that does not contain *w* or *qu*), resulting in nonwords like HACT (this condition is hereafter referred to as CS+BR+, as both the context-sensitive correspondences, CS, and the body-rime correspondences, BR, agree with the context-insensitive GPC *a* → "/æ/". The second condition (CS+BR-, e.g., HALSE) was based on bodies where the *a* is consistently pronounced as "/oː/" (or "/ɔ/" for the body *-att*), and "normal" onsets, as in the first condition. Here, the body-rime correspondences predict an "/oː/" pronunciation, and therefore disagree with the context-insensitive correspondence. The items in the third condition (CS-BR+, e.g., WACT) were based on regular bodies and onsets containing *w* or *qu*, meaning that the context-sensitive

[qu,w]a-correspondence contradicted the context-insensitive GPC while the body-rime correspondences did not. The fourth condition (CS-BR-, e.g., QUALSE) had items with irregular bodies and onsets with *w* or *qu* - here both the context-sensitive correspondence and the body disagree with the context-insensitive GPC. As filler items, we used a set of unrelated nonwords.

The presentation was identical to Experiment 1, with items presented in random order and in upper case letters. As with Experiment 1, participants were instructed to read the items as accurately as possible, without putting them under time pressure.

4.4.2. Results

The results were scored by the fourth author (SP), a native Australian English speaker and an experienced transcriber, with the aid of spectral analysis using the EMU speech database System and associated speech analysis tools (Cassidy & Harrington, 2001). SP was unaware of the aims of the experiment while she was transcribing the data. Unlike the German data, scoring the responses as correct or incorrect was more complicated. For the letter *a*, there are at least five plausible pronunciations: as in "cat", as in "false", as in "what", as in "cake", and as in "car". We considered only the first three responses, as they were predicted either by the context-insensitive GPC, *a* → "/æ/", the context-sensitive GPC, *[qu,w]a* → /ɔ/, or the body-rime correspondence *a[l]* → "/oː/" context-sensitive correspondence. Other responses and errors made up 4.09% of the CS+BR+ condition, 24.85% of the CS+BR- condition, 6.43% of the CS-BR+ condition, and 20.76% of the CS-BR- condition, and were excluded from the subsequent analyses. The percentage of "other" responses is particularly high for the BR- conditions, partly because in English, a post-vocalic *l* creates ambiguity in the pronunciation of the vowel, such that a long "/oː/" may become indistinguishable from the phoneme "/əʊ/".

Table 4.

Summary of vowel responses of the English (monolinguals 2A bilinguals 2B), predictions from the three types of correspondences (context-independent GPCs; context-sensitive correspondences; body-rime correspondences) and predictions from the model using the weights in Table 2.

	Responses	CS-BR- ("qualk")	CS+BR+ ("hangst")	CS-BR+ ("quadge")	CS+BR- ("hald")
Participant responses					
2A	%æ	8.12	96.20	76.04	39.18
	%ɔ	60.25	0.00	17.19	27.19
	%oː	10.63	0.00	0.88	8.77
2B	%æ	8.07	83.33	62.50	41.20
	%ɔ	38.24	0.93	19.91	27.31
	%oː	52.19	0.00	0.00	7.87
Correspondence predictions					
GPC	P(æ|GPC)	72.00	72.00	72.00	72.00
	P(ɔ|GPC)	5.00	5.00	5.00	5.00
	P(oː|GPC)	6.00	6.00	6.00	6.00
CSC	P(æ|CSC)	29.00	77.00	29.00	77.00
	P(ɔ|CSC)	47.00	0.00	47.00	0.00
	P(oː|CSC)	0.00	100.00	0.00	100.00
BRC	P(æ|BRC)	0.00	100.00	100.00	0.00
	P(ɔ|BRC)	0.00	0.00	0.00	0.00
	P(oː|BRC)	100.000	0.00	0.00	100.00
Model predictions					
2A	%æ	22.46	82.98	49.07	56.37
	%ɔ	33.35	0.14	33.35	0.14
	%oː	26.77	0.16	0.16	26.77
2B	%æ	18.88	85.34	55.39	48.84
	%ɔ	29.38	0.05	29.38	0.05
	%oː	36.57	0.07	0.07	36.57

The percentages of "/æ/", "/oː/" and "/ɔ/" responses are presented in Table 4, with the results from Experiment 2B for comparison.

4.4.2.1. Modelling Vowel Pronunciations in English

The modelling strategy for Experiment 2A and 2B required a small modification from that employed in Experiments 1A and 1B. In German, there are only two available vowel pronunciations for *a*: short and long. In Australian English, there are three pronunciations available for items of Experiment 2. This means that we now need three equations per item:

$$\begin{aligned} P(\text{æ}) &= \beta_{gpc} \times GPC_{\text{æ}} + \beta_{csc} \times CSC_{\text{æ}} + \beta_{brc} \times BRC_{\text{æ}} \\ P(\text{ɔ}) &= \beta_{gpc} \times GPC_{\text{ɔ}} + \beta_{csc} \times CSC_{\text{ɔ}} + \beta_{brc} \times BRC_{\text{ɔ}} \\ P(\text{oː}) &= \beta_{gpc} \times GPC_{\text{oː}} + \beta_{csc} \times CSC_{\text{oː}} + \beta_{brc} \times BRC_{\text{oː}} \\ & where\ \beta_j \in [0,1]\ and\ \Sigma\,\beta_j = 1 \end{aligned} \quad (5)$$

where each of the subscripted strategies indicates the likelihood of the subscripted pronunciation under that strategy; for example, $GPC_{\text{æ}}$ indicates the likelihood of an "/æ/" response under the GPC strategy. The end result is a set of $\hat{\beta}_i$s that fit all three pronunciations simultaneously.

The weightings are shown in Table 2. The role of context-sensitive correspondences appears to be the most important in predicting the pronunciation of the grapheme *a*, with, β_{csc} = 0.69. Body-rime correspondences also appear to contribute significantly, β_{brc} = 0.26, while the reliance on context-insensitive correspondences is very small, β_{gpc}= 0.05. Indeed, the bootstrapping procedure produced β_{gpc} = 0 in 43.3% of the samples, and β_{gpc} < 0.1 in 82.0%, suggesting that the reliance on context-insensitive correspondences does not differ significantly from zero. Here again, the model is outperforming each of the independent strategies at predicting response patterns on an item by item basis (see Table 3), but when considering the model's ability to predict cell

means (Table 4), it's clear this approach is less successful in English than it was in German.

4.4.3. Discussion

We quantified the reliance on different types of correspondences for English nonwords with the grapheme *a*, using the same modelling technique we introduced in Experiment 1 for German, with some minor modifications. Although the results were less clear-cut than in German, we show that the procedure can be applied to a more complex orthography. The model fits in Table 4 indicate that the English orthography is not best suited for such an analysis. In particular, the poor model fits are due to many "/ɔ/"-responses, even when these were not predicted by the model. This may be a result of the complex phonology of English: the phonemes "/ɔ/" and "/oː/" are very similar, therefore it is possible that the participants had a tendency to shorten "/oː/"-responses, which then became indistinguishable from the vowel "/ɔ/". The second possibility is that another source of information is used to determine vowel pronunciations in English which we did not take into account.

Despite these limitations, there are several conclusions that can be drawn from the results. Firstly, the weightings showed that in English the three strategies are neither necessary nor sufficient to predict the pronunciation of the grapheme *a*. In contrast to German, we obtained a relatively high percentage of "other" responses for the English data, or pronunciations that were implausible according to any of the correspondences that we thought participants may use. Such a heterogeneity of nonword reading aloud responses has also been reported elsewhere (Andrews & Scarratt, 1998; Pritchard et al., 2012). While this would be an interesting topic to pursue in further research, for our purposes we discarded the unusual pronunciations as we were interested in quantifying the reliance on the same three types of correspondences we showed to be critical to

nonword reading in German. This high percentage of "other" responses shows that it is likely that other strategies, such as more complex context-sensitive correspondences or lexical analogy, are used during nonword reading in English. In other words, the three types of correspondences we described in the introduction are not sufficient to explain vowel responses to the grapheme *a* in English - which is in contrast to the findings we report for German.

Secondly, a striking finding is that the context-insensitive correspondences are hardly used at all to derive the pronunciation of the grapheme *a*. Rather, English readers rely heavily on the context-sensitive GPC, which can often be used to derive the correct pronunciation for English words.

These results imply that in the special case of the grapheme *a*, it may not be necessary to rely on all three types of sublexical correspondences to explain the pattern of vowel responses. We consider it highly unlikely that context-insensitive GPCs are not used at all for reading in English. We relied solely on nonwords with the grapheme *a* to derive the weightings in Experiment 2, and its correct pronunciation can often be predicted by context. Arguably, this may falsely bias the weightings towards an apparent greater reliance on context-sensitive correspondences than we would observe if we used different graphemes for this procedure. However, we consider it likely that context-sensitivity plays an equally important role for other vowels in English: as is the case for the grapheme *a*, vowel pronunciations in English are generally inconsistent, but can be often resolved by context-sensitive correspondences (Treiman et al., 1995). Nonword reading studies have also provided evidence for the psychological reality of context-sensitive correspondences determining vowel pronunciation in English, other than the *[qu/w]a*-correspondence (Treiman et al., 2003; Treiman et al., 2006). As described above, we focussed on the *[qu/w]a*-correspondence only because it is not confounded with body-

rime correspondences - if we used any other context-sensitive correspondence we would be unable to distinguish it from reliance on body analogy.

Again, we stress that the almost exclusive reliance on context-sensitive correspondences in Experiment 2 is unlikely to generalise to the processing of more consistent graphemes in English, such as consonants. If, linguistically, context-insensitive correspondences are generally predictive of the correct pronunciation, there is no pressure on the readers to take into account the surrounding letters for those particular graphemes.

As discussed in the introduction, the body-rime correspondences of English are confounded with context-sensitive correspondences. Instead of the German super-rules, we used an English context-sensitive correspondence that is not located in the body, namely the *[qu,w]a* → "/ɔ/" correspondence. However, we cannot fully disambiguate the reliance on body-rime correspondences and the *a[l]*-correspondence. Future studies using nonword reading should bear in mind that body-rime correspondences and context-sensitive correspondences are heavily confounded, and that an apparently irregular pronunciation of a nonword may show reliance on either context-sensitive correspondences or body-rime correspondences.

4.5. Experiment 2B

In Experiment 2B, we tested a sample of German/English bilingual speakers on the English item set. As with Experiment 1B, this will allow us to verify the weightings in a different sample, and explore potential differences between mono- and bilingual participants.

In Experiment 1, we argued that the differences that we found between the two samples are more consistent with an account based on reading proficiency rather than one based on the influence of acquiring a language with a deeper orthography. However, it may be that an early acquired L1 shapes the cognitive system in a way that biases the

processing of subsequently learnt languages towards familiar types of correspondences. If so, this would predict a difference between participants reading English nonwords depending on whether their first language was English (as in Experiment 2A) or German.

4.5.1. Methods

The participants were 13 native German speakers living in Australia (undergraduate and graduate students at Macquarie University, academic staff, family and friends). Eight of them had also participated in Experiment 1A several months earlier, but did not know that the two studies were related. In this sample, all participants had lived in Germany for at least 18 years before moving to an English-speaking country. The items and procedure were identical to Experiment 2A. The participants were told that they would see English nonwords, and were asked to pronounce each item as if it were an English word that they are unfamiliar with.

4.5.2. Results

The same scoring system was used as for Experiment 2A. The proportions of "/æ/", "/ɔ/" and "/oː/" responses for both Experiment 2A and 2B are presented in Table 4. German native speakers overall gave more "other" nonword responses, or vowel responses that were inconsistent with our predictors, compared to the English monolinguals in Experiment 2A: 15.74%, 23.61%, 17.95%, and 8.80% for the CS+BR+, CS+BR-, CS-BR+ and CS-BR- conditions respectively.

We repeated the optimisation technique to derive strategy weights for this Experiment. Table 2 summarises the weights for each of the three strategies in Experiments 1A, 1B, 2A and 2B. The results of Experiment 2B mirror the findings from Experiment 2A: Again, we find strongest reliance on context-sensitive correspondences, robust reliance on body-rime correspondences, and negligible reliance on context-insensitive correspondences. Numerically, the reliance on context-sensitive

correspondences appear to be larger ($\hat{\beta}_{csc}$ = 0.61) than in the monolingual sample ($\hat{\beta}_{csc}$ = 0.69). Here again, the optimal parameters outperform the alternatives with a correlation of .717 (see Table 3).

4.5.2.1. Comparing bilingual to monolingual English readers

Using the same bootstrapping technique described in Experiment 1, we confirmed that the German-English bilingual participants relied more on body-rime correspondences (BRCs) than did the English monolinguals. In 9,998 (99.98%) of the samples, $\hat{\beta}_{brc}$ was larger for bilinguals than monolinguals (95%CI of the difference: 0.046 to 0.150). The two samples did not differ significantly in their reliance on context-insensitive (GPC) rules, but there is some evidence that the monolinguals may rely more on context-sensitive correspondences (91.72% of the samples, 95%CI: -0.039 to 0.160).

4.5.3. Discussion

In Experiment 2B we collected data on English nonword pronunciation from German/English bilingual participants, which we then compared to the *a*-pronunciations of English monolinguals in Experiment 2A. Again, we find that the fits of the model are somewhat discrepant with the data, suggesting that the pronunciation of the letter *a* depends also on sources of information that are not included in our model. As in Experiment 2A, we found no reliance on context-insensitive GPCs in either group, and only a nonsignificant trend towards larger reliance on body-rime correspondences or the *a[l]* → "/oː/" correspondence in English monolinguals than the German/English bilinguals.

We found broadly the same pattern among two different groups of participants; here, we once again demonstrate the reliability of the optimisation procedure. The significant difference in the reliance on body-rime correspondences suggest that German

native speakers, when they are highly proficient in English, rely more on these large units than English monolingual participants. Thus, the native orthography does not appear to leave footprints in the cognitive processes underlying reading in a second language, as in this case we would expect diminished reliance on BRC in German compared to English native speakers.

4.6. General discussion

In four experiments, we explored the reliance on three different sublexical correspondence types in different populations. In Experiments 1A and 1B, we found that German native speakers relied on all three strategies: the greatest weighting was found for context-insensitive GPCs, followed by context-sensitive GPCs (super-rules) and body-rime correspondences when reading German-derived nonwords. In Experiments 2A and 2B, we applied the same procedure to quantify the types of correspondences that participants rely on to derive the pronunciation of the grapheme *a* in English. We found strong reliance on context-sensitive GPCs, some reliance on body-rime correspondences, and little evidence that context-insensitive GPCs play a large role in determining the pronunciation of the grapheme *a*.

4.6.1. Cross-Linguistic Differences in the Choice of Sublexical Correspondences: Comparing Experiments 1 and 2

Previous theoretical work predicts cross-linguistic differences in the reliance on different units in German and English (Ziegler & Goswami, 2005). Unfortunately, with the experiments in the current study it is impossible to make a direct quantitative comparison across the two languages as we are comparing two differently structured orthographic correspondences. An alternative approach is to conduct the analyses within the languages and point out the differences between them on a descriptive level.

Our data suggest that given a grapheme where context is very important in English (i.e., *a*), context-sensitivity becomes very important compared to German, where context-insensitive correspondences are the major predictor. This is true even for a situation where there are statistical regularities at the level of context-sensitive correspondences. This is broadly in line with the psycholinguistic grain size theory (Ziegler & Goswami, 2005): as the context is often an important predictor of the correct pronunciation of English words, readers are forced to rely on larger units. Our data emphasises the importance of context-sensitive GPCs in an inconsistent orthography such as English. In German, on the other hand, context-insensitive correspondences are mostly sufficient to derive the correct pronunciation of an unfamiliar word, therefore this level of correspondences is preferred.

The reality of the cross-linguistic differences becomes more evident in a comparison of Experiments 1A and 2B. This is partly a within-subject design, and involves bilingual participants reading both the English and the German item sets. The differences between the weightings in these two experiments were remarkable, with the pattern of results being more similar to that of the monolinguals of the respective language. This shows that the language is the determining factor for the reliance on different unit sizes, rather than the language background of the participants.

From this comparison, we conclude that the language that a participant is asked to read in matters more than the participant's language background: comparing the participants in Experiments 1A and 2B shows that bilinguals rely on the three types of correspondences almost to the same extent as monolinguals do in their respective language. Thus, we conclude that the cross-linguistic differences in sublexical processing are language-specific: acquiring a deep versus shallow orthography from childhood does not shape the cognitive system, but rather encourages the reader to rely on certain types

of correspondences above others in that particular orthography. Those preferences do not seem directly transferable to a later acquired orthography; instead, a reader develops a sensitivity to the most advantageous combination of strategies in the new language.

4.6.2. Models of reading

The current study shows that both in English and in German, several correspondence types are used in parallel. There are multiple verbal models that postulate such a scenario (LaBerge & Samuels, 1974; Patterson & Morton, 1985; Taft, 1991; Ziegler & Goswami, 2005). The theoretical contribution of the current paper is proposing a method to quantify the degree to which these are used, which can be used as a benchmark for computational models.

An open question then is whether the current computational models can simulate the obtained results. The parallel processing of various correspondences poses a computational problem: whenever there are conflicts between the pronunciations predicted by various correspondences, the system needs a way to resolve these. In English, this is important, because there are often cases where different sublexical correspondences provide conflicting information.

In Table 5, we provide the percentages of regular responses from two models which have been implemented both in English and in German, namely the DRC (M. Coltheart et al., 2001; Ziegler et al., 2000) and the CDP+ (Perry, Ziegler, Braun, et al., 2010; Perry et al., 2007). For English, there is a newer version of the CDP+, namely the CDP++ (Perry, Ziegler, & Zorzi, 2010), which differs from the CDP+ in several points: it has been trained on a larger word set, contains some parameter changes, and can also deal with polysyllabic words. We provide the simulation data from both versions of the model.

Both the CDP+/CDP++ and the DRC are dual route models of reading, where nonwords are read purely via a sublexical procedure. Therefore, the current data are

relevant to both models, as it concerns the nature of sublexical processing. The distinguishing feature between the two models is the way in which this procedure operates. The DRC has a set of sublexical GPCs, which are manually programmed into the sublexical route. A GPC in the DRC is defined as the most frequent phoneme that co-occurs with a given grapheme. As described in the introduction, the DRC contains context-sensitive correspondences as well as single-letter and multi-letter correspondences, but there is some ambiguity when it comes to deciding which context-sensitive correspondences to include in the model. The current version of the English DRC does not contain either a *[w]a-* or an *a[l]*-correspondence, therefore it provides the response "/æ/" to all items (see Table 5).

Table 5.

Percentage of "regular" responses ("/æ/" in English, short vowels in German) given by the DRC and CDP+/CDP++

	CS+BR+	CS-BR+	CS+BR-	CS-BR-
English				
Behavioural	100	81	51	18
DRC (sim. 1)	100	100	100	100
DRC (sim. 2)	100	67	11	0
CDP+	100	35	57	0
CDP++	100	73	44	0
German				
Behavioural	86	73	63	37
DRC	100	0	100	0
CDP+	93	94	8	24

For the second DRC simulation, we added some more context-sensitive correspondences, however this does not seem to reflect the overall responses given by participants, either, as it now underestimates the number of regular (i.e., "/æ/") pronunciations given by the participants. For the German DRC, the GPCs that are used to

determine vowel length are the super-rules (Ziegler et al., 2000). It is clear, both from the present study (see Table 5) and from Perry et al. (2010) that the super-rules are not sufficient to explain German nonword pronunciations.

The CDP+/CDP++, like the DRC, is grapheme-based, but it develops context-sensitivity because the grapheme-to-phoneme correspondences are derived via a learning algorithm, which uses real word knowledge to obtain the most likely correspondences between print and speech (Zorzi, 2010). Yet, the CDP+ does not provide an optimal fit for either the German or the English data, as it often underestimates the number of regular pronunciations (see Table 5). In particular, the English CDP+ and CDP++ seem to take context-sensitive correspondences into account more than the participants do, as they underestimate the number of "/æ/"-responses for the CS- conditions. In German, the biggest discrepancy between the CDP+ prediction and the behavioural data is in the BR-conditions, suggesting that CDP+ does not develop the same degree of reliance on body-rime correspondences that participants do.

As neither of the computational models is compatible with the behavioural results, these data cannot be used to adjudicate between the DRC and CDP+ approach. (Note that this was not the aim of the study to begin with.) We therefore turn to verbal models to provide a theoretical framework that can explain our obtained results. One such model which provides a means for the cognitive system to resolve conflicts between different sublexical correspondences has been proposed by Taft (1991). This interactive activation model states that activation passes hierarchically from the smallest units, through subsyllabic and syllabic units and morphemes to whole words, which then gives access to the semantic concept. There are additional feedback connections, which send activation from larger to smaller units.

Taft's (1991) model also makes some explicit statements about cross-linguistic differences: the salient sublexical correspondences differ depending on the orthographic and phonological properties of the language. For example, while English readers parse words into orthographic-syllabic units called BOSSes (Taft, 1979, 1992) French readers rely more on the phonological syllable (Taft & Radeau, 1995). In our experiments we found reliance on similar types of correspondences in English and German. Thus, the correspondences that have psychological reality in English and German appear to be very similar. It is noteworthy that English and German are very similar in terms of their phonological and orthographic structure, therefore we expect that the salient sublexical correspondences do not differ greatly. The situation might be different in other languages. For example, when there is a tendency for words to be polysyllabic and to contain fewer consonant clusters, as is the case in languages like Italian, Spanish, or Russian, body-rime correspondences are unlikely to play a large role in reading (Duncan et al., 2013; Kerek & Niemi, 2012).

4.6.3. Limitations and future directions

The goal of the study was to identify an optimal combination of different sources of information in deciding which vowel pronunciation is most appropriate when there are two or more alternatives. A limitation of the model is that it makes no claims about the decision-making mechanisms that resolve the ambiguity, only that some sources of information are more influential than others. It may be that on each trial, the decision is based on a "winning strategy" in which case the weights represent the likelihood of a particular strategy winning. Alternately, it may be that all three sources of information are combined in a Bayesian sense of "what response is most likely correct given the mix of influences." In this case the model weights should be interpreted as the degree of influence that each strategy has on the decision process. The present study is not able to

adjudicate between these alternatives (or any others that we may not have considered), so we refrain from making strong statements favouring one or the other. The extent to which nonword pronunciations remain stable in different situations, the factors that influence any variability, and the mechanisms that resolve ambiguity remain questions for future research. We do note, however that while there is considerably variability between subjects in terms of their strategy weights (see Figure 1), there is some recent evidence that readers can be grouped according to their choices (Robidoux & Pritchard, 2014), so there may be more structure hiding within this variability.

A limitation of the paradigm as described in this paper is that it is better suited for across-subject comparisons than across-item comparisons, due to the small number of available items. This is a general problem with this approach: there are not many items where context-sensitive correspondences and body-rime correspondences can be dissociated, as these are intrinsically correlated. While it would be interesting to use the same paradigm for a different set of nonword or word items to explore systematic changes in the weightings associated with item characteristics such as frequency (for words) or word-likeness (as measured, e.g., by orthographic N), the small number of possible items prevents us from doing this in a meaningful way.

Arguably, the data reported in this paper are also limited by our focus on the grapheme *a* only. While this criticism applies to the English data, the German data can be generalised to predicting vowel length across different graphemes. The English results, and our conclusions based on these analyses, are therefore weaker than those from the German analyses. Nevertheless, understanding the principles underlying reading in languages other than English is essential for the long-term goal of describing all differences and similarities between reading in different languages, and thereby creating a universal model of reading (Frost, 2012a). This is especially important given the focus of

previous literature on English. English is considered to be an "outlier" orthography, therefore it is questionable to use it as a base for most models of skilled reading, reading development, and dyslexia (Share, 2008). Although we acknowledge that, in the current context, the optimisation procedure works better for German than English, we argue that the English data provides a strong demonstration of the parallel use of different types of sublexical grain sizes, and in particular context-sensitive correspondences in English, new insights into cross-linguistic differences associated with the reliability of print-to-speech correspondences, and a new benchmark for computational models of reading aloud.

We believe that this approach also has some utility when applied to other areas of psycholinguistics. In future research, the same paradigm can be used to systematically explore the sources of individual differences that we report in the current study. The paradigm can also be used with children: previous literature has debated for decades whether children start learning to read using large or small units first (Goswami, 2002; Goswami & Bryant, 1990; Hulme et al., 2002). Such explorations in group and individual differences are of theoretical and practical value. Future research can also apply the same mathematical procedure to any situation in which items can be created where different strategies yield different predictions. Other areas in psycholinguistics to which this paradigm can be extended could be topics such as stress assignment for polysyllabic words, because it has been shown that, in several languages, different cues are used by participants to determine the stress of a given nonword (Arciuli, Monaghan, & Seva, 2010; C. Burani & L. S. Arduino, 2004; Protopapas, Gerakaki, & Alexandri, 2006; Seva, Monaghan, & Arciuli, 2009).

4.7. Conclusions

The current study contributes to the literature on cognitive processes underlying reading in several aspects. We show that context-insensitive GPCs, super-rules and body-

rime correspondences are necessary and sufficient to explain the vowel length pronunciations in German; in English, context-insensitive GPCs play a smaller or negligible role in assigning the pronunciation of the grapheme *a*. We introduce a method to quantify the degree of reliance on each of the three different sublexical correspondence types using statistical modelling. This technique can be used to test other hypotheses by future studies.

Appendix A: German and English nonwords used in Experiments 1 and 2

German items

V[C] Reg	*V[C][C] Reg*	*V[C] Irreg*	*V[C][C] Irreg*
blaf	bamt	bax	bags
blen (*blem)*	birt	blex	blags
blod	blaft	blig	füst
breg	bling	bres	gleks
brel	boft	flim	kagd
brul	brals	flis	kagt
flom	chrolf	git	kets
flüb	falb	glef	pagt
fryp	flarg	glip	pard
grät	flerk	krex	peks
grem	gärm	krin	poks
grom	ginn	krip	schagd
grul	gralb	pfis	stard
klid	gunt	spic	
klur	kall	stef	
knul	kaxt	zwix	
krel	kerv	zwok	
kril	kluns		
krön	knell		
pflyp	pals		
pid	peld		
plät	pfern		
plön	pulk		
prod	purf		
schmün	schern		
schraf	spalf		
schwüb	stelf		
speg	sturg		
zwül	zeng		

zwun	zwurt

English items

CS+BR+	*CS+BR-*	*CS-BR+*	*CS-BR-*
hangst	clatt	quadge	qualk
kazz	hald	quamb	qualse
mact	halse	quangst	qualtz
phadge	kalk	quapse	squald
phamb	kalse	quazz	squalk
phangst	kalt	squact	squalse
phants	phalk	squazz	squaltz
plact	phaltz	swact	swalk
sangst	slaltz	swangst	swaltz
slangs	strald	swants	twald
slazz	stralk	swazz	twalk
stract	stralse	twadge	twalse
stramb	straltz	twangst	twalt
tamb	tald	twants	twaltz
tazz	taltz	twazz	wald
tradge	tralse	wact	walse
trazz	tralt	wamb	walt
zants	tratt	wangst	whald

Appendix B: Implementing the fitting in R

While fitting the models described in the text has a certain flavour of regression to it, there are some important differences. Most critically are the two constraints that we have placed on the parameters: $\beta_j \in [0,1]$ and $\Sigma\, \beta_j = 1$. Considerable work has been done to develop and implement estimation methods for models with inequality constraints such as $\beta_j \in [0,1]$ (Grömping, 2010). However, we know of no such work that has solved the problems presented by the $\Sigma\, \beta_j = 1$ constraint. To address this problem, we turned to the *optim* function that is part of the base statistical analysis package in R (R Core Team, 2013). *Optim* is a very general optimisation package that allows the user to minimise any specified function, while also placing bounds on the returned values. That is, we can define a function, place upper and lower bounds on the returned weights, and *optim* will efficiently search the allowed parameter space to minimise our function. To satisfy $\beta_j \in [0,1]$, we defined the minimising function to be the residual sum of squares, and restricted the β weights to fall between 0 and 1. This ensures that $\hat{b}_j \in [0,1]$ is satisfied.

In all of the optimisation analyses, we used the following command in R: *optim(par=runif(3, .2, .8), fn=..., ..., method='L-BFGS-B', control=list(factr=1e5), lower=0, upper=1)}.* The parameters for optim operate as follows: "*par=runif(3, .2, .8)"* initialises the β_i's to random values between .2 and .8. "*fn=..., ...*" specifies the function to be minimised along with any parameters it requires. In our case we used a simple function that calculates the residual sum of squares. *"method='L-BFGS-B'"* instructs *optim* to use an optimisation algorithm that allows for upper and lower bounds on the returned values (Byrd, Lu, Nocedal, & Zhu, 1995). "*factr=1e5*" sets the convergence tolerance, and "*lower=0, upper=1*" set the bounds on the returned values.

$\Sigma\, \beta_j = 1$ *constraint*

There is no way to explicitly tell *optim* to meet the constraint that the βs must sum to 1 ($\Sigma\, \beta_j = 1$). One way to ensure that the constraint is met is to simply scale the weights returned by *optim* using the formula:

$$\beta_j' = {\beta_j}\Big/{\Sigma\beta_j}$$

where β_j' are the new scaled weights, and are guaranteed to sum to 1. However, since this adjustment *follows* the optimisation process, there is little reason to believe that the resulting β_j's would remain an optimal solution to (2). An alternative to simply scaling the β_j's, is to make use of the influence of outliers on parameter estimation. For example, according to (2) *optim* is trying to satisfy the following 180 equations (two per item) simultaneously, by minimising the residual sum of squares (while also meeting the $\beta_j \in [0,1]$ constraint):

$$P_1(Short) = \beta_{gpc} \times GPC_{short,1} + \beta_{csc} \times CSC_{short,1} + \beta_{brc} \times BRC_{short,1}$$

$$P_1(Long) = \beta_{gpc} \times GPC_{long,1} + \beta_{csc} \times CSC_{long,1} + \beta_{brc} \times BRC_{long,1}$$

$$\ldots \qquad (6)$$

$$P_{90}(Short) = \beta_{gpc} \times GPC_{short,90} + \beta_{csc} \times CSC_{short,90} + \beta_{brc} \times BRC_{short,90}$$

$$P_{90}(Long) = \beta_{gpc} \times GPC_{long,90} + \beta_{csc} \times CSC_{long,90} + \beta_{brc} \times BRC_{long,90}$$

The introduction of a new data point that can only be met by satisfying the constraint that the $\Sigma\, \beta_j = 1$ will put some pressure on *optim* to select appropriate parameters. For example,

$$1 = \beta_{gpc} \times 1 + \beta_{csc} \times 1 + \beta_{brc} \times 1 \qquad (7)$$

Equation 7 is equivalent to creating an artificial data point where all of the dependent and independent variables [P(Short), GPC, CSC, and BRC] are set to 1. Though (7) provides some pressure to satisfy $\Sigma\ \hat{b}_j = 1$, it is unlikely to have a very large influence since it is

only a single equation with roughly equal weight to the other 180. However, dramatically increasing the weight of this data point will exert a much stronger influence on the final parameter selection. For example,

$$10{,}000 = \beta_{gpc} \times 10{,}000 + \beta_{csc} \times 10{,}000 + \beta_{brc} \times 10{,}000 \quad (8)$$

Equation 8 would put enormous pressure on *optim* to arrive at a set of weights that satisfy $\Sigma\ \hat{b}_j = 1$,without putting any further constraints on how the weights are apportioned to the strategies. Though Equation (8) does not guarantee $\Sigma\ \hat{b}_j = 1$, precisely, it is sufficiently strong for the present purposes. Other applications may require a larger multiplier.

Finally, because the number of items is not equal across all conditions in our studies, the sums of squares were weighted by item to ensure each condition contributed equally. For example in Experiment 1, items in the V[C] Irregular and V[C][C] Irregular conditions received relatively more weight than items in the V[C] Regular and V[C][C] Regular conditions. If this isn't done, there is a tendency for the Regular items to have a stronger influence on the eventual parameters. The weights applied to each item were determined as follows:

$$\omega_{type} = \frac{.25}{n_{type}}$$

where *type* is one of the four item types (e.g., V[C][C] Irregular in Experiment 1), ω_{type} is the weight assigned to items of that type, and n_{type} is the total number of items of that type. As this formula implies, each item contributes equally to the influence of its category, but items in smaller categories have more influence than items in larger categories. These weights are then used in the usual weighted sum of squares formula that optim is trying to minimise:

$$SS_{resid} = \sum_i \left(\hat{Y}_i - Y_i\right)^2 \omega_{type_i}.$$

Paper 5: Lexical and sublexical processing in English and German children

Abstract

Previous research has shown that in orthographies with complex and unreliable sublexical correspondences (e.g., English), reading acquisition takes longer than in orthographies where such correspondences are more transparent (e.g., German). Here, we examine cross-linguistic differences in the efficiency of lexical and sublexical processing in English and German children. This is indexed by irregular and nonword reading respectively. We also examine the nature of sublexical processing across grades. The first question we addressed is whether unreliability of sublexical information impedes the acquisition of lexical processing. This appears to be the case, as both nonword and irregular word reading were poorer in English than in German. Secondly, we explored differences across age in reliance on sublexical correspondences by using a nonword reading paradigm in conjunction with an optimisation procedure, where we use the participants' nonword pronunciations to deduce what types of sublexical units were used to derive it. Overall, the results from the optimisation are more reliable in German than English. Children in both languages become more sensitive to more complex units with age. Complex correspondences are used to a greater extent in English compared to German.

5.1. Lexical and sublexical processing in German and English children

Reading is an important skill in everyday life, and a large amount of research has been dedicated to understanding how this highly complex skill is acquired. A large proportion of theories of reading acquisition, however, are based on studies which have been conducted in English-speaking countries (Share, 2008). Although recent work has been conducted to amend this situation (e.g., see Ziegler & Goswami, 2005), open questions remain about differences and similarities in cognitive processes underlying reading acquisition across languages. Specifically, the English orthography is notorious for the complexity of its print-to-speech translation system compared to other European orthographies. This has been consistently shown to impact the speed of reading acquisition across European orthographies (e.g., Seymour et al., 2003). Yet, the exact cognitive mechanisms that are affected by this complexity - or, in fact, what specific language features differ across languages and produce the effects - are still unclear (see Paper 6).

In the current study, we investigated in detail the developmental trajectory of cognitive processes underlying reading acquisition in English and German. These two languages form an interesting comparison: due to their common Germanic origin, they are similar in terms of orthographic and phonological structure, but differ in terms of orthographic depth, or the transparency of the principles that underlie print-to-speech conversion: German correspondences are relatively straight-forward ("shallow"), while English relies on more complex and less transparent ("deep") principles (Frith et al., 1998; Landerl et al., 1997; Wimmer & Goswami, 1994; Ziegler, Perry, Ma-Wyatt, et al., 2003).

Orthographic depth, and the effect that it may have on reading, has been studied for decades (see Katz & Frost, 1992; Ziegler & Goswami, 2005, and Paper 6, for

reviews). It is well established that learning to read in a deep orthography takes longer compared to learning to read in a shallow orthography (Frith et al., 1998; Seymour et al., 2003). Empirical studies unanimously report this finding, even when socio-cultural differences and differences in reading instruction are controlled for (Bruck et al., 1997; Caravolas & Bruck, 1993; Ellis & Hooper, 2001; Landerl, 2000).

Here, we aimed to address two questions: (1) whether, for children learning to read, both lexical and sublexical processing skills are harder to acquire in English than in German, and (2) how the nature of sublexical processing differs across languages and across development. The first question relates to a distinction between lexical and sublexical processes during reading. Most models of reading contain a division between these two procedures (M. Coltheart et al., 2001; Perry et al., 2007; but see Plaut, 1999; Plaut et al., 1996, for an alternative framework). A sublexical procedure uses correspondences smaller than words to assemble the pronunciation of a written input, and can therefore be used for unfamiliar words. Whole-word knowledge of a real, familiar written word allows for direct access to its pronunciation and meaning. The store of orthographic word forms and their corresponding links make up the lexical route.

As orthographic depth relates to the reliability of *sublexical* correspondences, it is an intuitive prediction that learning to use this procedure efficiently would be more difficult in deep compared to shallow orthographies. This has also been confirmed through a considerable number of studies which have compared nonword reading accuracy across orthographies varying in depth (e.g., Frith et al., 1998; Landerl, 2000; Landerl et al., 1997; Seymour et al., 2003). The importance of sublexical processing for reading acquisition has been consistently demonstrated (for reviews, see Share, 1995; Ziegler & Goswami, 2005), therefore this is likely to be a major source of the overall lag in reading acquisition of English compared to German.

It is less straightforward to predict how orthographic depth might impact on the development of lexical processing, but based on existing theories of reading we can make some predictions, which we aimed to test in the current study. According to the self-teaching hypothesis (Share, 1995), the sublexical procedure is essential for the development of an efficient lexical route, as it acts as a mechanism that allows a child to decode unfamiliar written words and match them to a familiar spoken word form. This is especially important for younger children, as in the beginning of reading acquisition all orthographic word forms are unfamiliar. This matching between written word forms and familiar spoken word forms is postulated to establish a connection between the two, which enables the reader to build up orthographic entries in the mental lexicon. Taking into account that (a) sublexical processing is harder to learn in English than in German, and (b) sublexical processing is necessary in order to establish a sound orthographic lexicon according to this theory, we can hypothesise that not only the acquisition of efficient sublexical, but also that of a sound lexical processes will be impaired in English compared to German. Our first aim was to test this prediction.

The second aim of the current study was to explore the developmental trajectories of the specific sublexical mechanisms underlying reading across orthographies. In particular, sublexical processing can occur via different types of correspondences, such as graphemes (letters or letter clusters that correspond to phonemes, or the smallest units of speech, such as *t* or *th*), or bodies (for a monosyllabic word, this is the vowel and coda of the written word form, e.g., *-ord*, *-iend*). The phonological equivalent of a body is called the rime. In the reading acquisition literature, a major debate revolves around the types of units which best predict individual differences in reading acquisition (e.g., Caravolas et al., 2005; Duncan et al., 1997; Goswami, 2002; Goswami & Bryant, 1990; Hulme et al., 2002). These studies focus in particular on children's sensitivity to *phonological* units

(phonemes, rimes), as phonological awareness has been consistently shown as a strong predictor of individual differences in reading acquisition (e.g., see Goswami & Bryant, 1990; Ziegler & Goswami, 2005, for reviews). Previous studies suggest a complex pattern of phonological development which is interdependent with reading acquisition, spoken language characteristics, reading instruction methods, and the type of task that children are asked to perform (e.g., Brown & Deavers, 1999; see Duncan et al., 2013, for a recent large-scale cross-linguistic study).

In the context of reading, it is important to take this question to the next level: namely, what factors drive the development of reliance on different *orthographic* units (graphemes, bodies), and the connections to their phonological equivalents? This is an important theoretical link which is needed to bridge the gap between our knowledge of phonological development, and the well-established finding that phonological awareness is related to reading acquisition. According to the psycholinguistic grain size theory (Ziegler & Goswami, 2005), which is a major theory of reading acquisition across languages, the consistency of the overall match between orthography and phonology (i.e., orthographic depth) is a critical factor: if there is no one-to-one correspondence between letters and sounds, readers are forced to rely on larger orthographic units, which have a more consistent link to phonology (Ziegler & Goswami, 2005).

Hypotheses about cross-linguistic differences in the types of correspondences used by developing readers as a function of orthographic depth can be tested in English and German, using a nonword reading paradigm in conjunction with an optimisation procedure. This has been recently done for skilled adult readers (Schmalz et al., 2014). Here, sets of nonwords are created whose pronunciations will differ as a function of what type of sublexical correspondences are used. Schmalz et al. (2014) compared three types of sublexical correspondences: context-insensitive correspondences (in English, that *a* is

pronounced as in "cat"), context-sensitive correspondences (*a* is pronounced as in "swan" when preceded by a *qu* or *w*), and body-rime correspondences (words ending with the body *-ald* are pronounced as in "bald"). For example, for a nonword like *quald*, the pronunciation would be "/kwæld/" if participants rely on context-insensitive correspondences, "/kwɔld/" if participants rely on context-sensitive correspondences, and "/kwoːld/" if body-rime correspondences are used.

Based on the percentage of different vowel responses for each nonword, Schmalz et al. (2014) then ran a mathematical optimisation procedure to compute the relative influence of each of the three different types of correspondences for a given individual. Overall, the procedure worked better for German than English: the German results showed that the three types of correspondences are both necessary and sufficient to predict the participants' vowel responses.

For English, the three types of correspondences were neither necessary (as the overall reliance on context-insensitive rules was negligible) nor sufficient (as there were many vowel responses which were not predicted by either of the three correspondences). Some broad cross-linguistic differences emerged: in German, context-insensitive correspondences were the strongest predictor of vowel responses, while in English, context-sensitive correspondences were the strongest predictor. This is in line with the psycholinguistic grain size theory (Ziegler & Goswami, 2005), in the sense that it shows that English readers rely on complex correspondences to resolve inconsistencies in the letter-sound correspondences in their orthography, while German readers can rely to a greater extent on simple correspondences that do not take context into account.

Applying this nonword reading and optimisation procedure to developing readers can aid us in addressing the following questions: (1) Do we find the same cross-linguistic differences in children as in adults? (2) What is the developmental trajectory of the

emergence of any cross-linguistic differences in reliance on the correspondences? To date, it is not clear whether any cross-linguistic differences would emerge from the beginning of reading acquisition and decrease with reading experience, or whether the mechanisms underlying very early reading acquisition are independent of orthographic depth, and any differences would emerge with increased exposure to the orthography and its statistical characteristics.

In summary, based on previous theoretical and empirical work, we can make several testable predictions about cross-linguistic differences in cognitive processes during reading acquisition. Based on the self-teaching hypothesis (Share, 1995), we can make the following prediction: if sublexical processing is slower to develop in English due to the number and complexity of correspondences that need to be learnt, this should impede the build-up of entries in the orthographic lexicon, thus hampering the acquisition of lexical reading processes. Furthermore, the psycholinguistic grain size theory (Ziegler & Goswami, 2005) states that there are differences in the nature of sublexical processing across languages, such that English-speaking children should rely on larger sublexical correspondences than German children. It is largely unclear how the reliance on different types might unfold across orthographies. We set out to explore these questions, using two languages that differ in terms of their orthographic depth, namely English, a relatively deep orthography, and German, a relatively shallow one (Borgwaldt et al., 2005; Seymour et al., 2003).

5.2. Methods

5.2.1. Participants

The participants were 64 German-speaking children from Grades 2-4, and 62 English-speaking children from Grades 1-4. Their ages ranged from 7 years, 1 month to 10 years, 10 months (see Table 1). The German children had participated in an unrelated

study conducted by Potsdam University, and were invited to come back for another session. They were tested individually in a quiet room at Potsdam University. The English-speaking children were either recruited from two independent schools in rural New South Wales in Australia (N = 32), or participated in a large-scale developmental study conducted at Macquarie University (N = 30). All children were monolingual native speakers of their respective language. In both countries, testing took place in the second half of the school year. The children's ages and performances on various reading tests (described below) are summarised in Table 1.

5.2.2. Tests

5.2.2.1. Overall reading ability

To compare the children's overall reading efficiency, we developed a speeded word reading test based mainly on cognates, which we call the German/English Cognates (GEco) Test. Words that were not strictly cognates were matched on meaning, length, and subsyllabic structure (e.g., *zusammen - together*).

Table 1.

Participant characteristics; mean (SD). Note: GEco is number of words read correctly in 45 seconds from the German/English cognates test.

	English				German			
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 1	Grade 2	Grade 3	Grade 4
Number	9	25	17	11	0	24	19	21
Age	7;4 (0;2)	8;2 (0;4)	9;0 (0;5)	10;1 (0;3)	NA	8;0 (0;4)	9;1 (0;7)	9;11 (0;5)
Per-centile	85.04 (17.43)	69.51 (30.64)	62.74 (27.34)	49.25 (38.80)	NA	50.63 (28.37)	62.84 (29.03)	60.74 (28.86)
GEco*	51.67 (16.32)	55.76 (17.40)	58.71 (13.86)	54.89 (17.90)	NA	40.22 (12.75)	55.62 (11.62)	57.42 (12.48)

The test contained 100 items, which were printed in four columns on an A4 sheet of paper, and ordered by difficulty. Difficulty was determined by both length and frequency of the words. The children were given 45 seconds to read as many words as they could.

To ensure the validity of our test, and to provide standardised scores of the children's reading performance, we also used the TOWRE speeded word test for English (Marinus, Kohnen, & McArthur, 2013; Torgesen et al., 1999), and the one-minute speeded word reading task from the Salzburger Lese- und Rechtschreibtest II (SLRT) for German (Moll & Landerl, 2010). Both tests have a similar layout to the GEco test: for the TOWRE, the children are given 45 seconds to read as many words from the list as they can, and for the SLRT, they are given one minute. Indeed, the scores of the GEco and the two standardised reading tests were highly correlated, suggesting that they assess the same construct, both $r > 0.9$.

5.2.2.2. Lexical and sublexical processing

Traditionally, the efficiency of the sublexical and lexical route in English is assessed using nonwords and irregular words respectively (Castles & Coltheart, 1993). Nonwords cannot have an entry in the mental lexicon and have to be decoded sublexically. The number of correctly read nonwords, therefore, reflects the functioning of the sublexical route. Irregular words are words that do not comply to print-to-speech correspondence rules, such as *yacht* → "/jɔt/". Here, the lexical route is needed to derive the correct pronunciation, as sublexical decoding cannot provide a correct response and would output a regularisation error ("/jætʃt/").

We designed a set of nonwords matched as closely as possible across language for difficulty (length, orthographic N, and, for words, frequency). The German language has few words that break the letter-to-sound correspondences. Irregular German words tend to be either loanwords, or subtle irregularities that involve the unpredictable

pronunciation of vowel length (Ziegler et al., 2000). Since subtle irregularities may not be strong enough to provide a conflict between the correct and the rule-based pronunciation, we used loanwords as irregular German words (e.g., *Chef*, *Trainer*). The English irregular words were defined as irregular by the DRC (M. Coltheart et al., 2001). These irregularities could not be resolved by relying on larger correspondences.

We chose 17 irregular words that we judged to be familiar to children, at least in their spoken form, in each language. The English words were all listed in the Children's Printed Word Database (Masterson, Stuart, Dixon, Lovejoy, & Lovejoy, 2003), confirming that these words are likely to be known to children. No equivalent database is available for German, therefore familiarity was decided on the first author's judgement. Across languages, the words were matched for length, orthographic N, and frequency (Duyck et al., 2004). We then used Wuggy (Keuleers & Brysbaert, 2010) to generate nonwords to match each of the 17 irregular words of each language in terms of syllabic structure.

The final items (17 irregular words and 17 nonwords for each language) are listed in the appendix. They were printed on flashcards and laminated, and presented to each child in the same order, by increasing difficulty. The children were given an unlimited amount of time for each item, and were allowed to skip if the item was too difficult. Only accuracy was scored (as in Castles & Coltheart, 1993; Castles et al., 2009).

5.2.2.3. The nature of sublexical processing

As discussed in the introduction, we used the optimisation procedure introduced in Schmalz et al. (2014; i.e., Paper 4) to quantify the degree of reliance on context-insensitive correspondences, context-sensitive correspondences, and (in German) body-rime correspondences. This paradigm uses four different conditions of nonwords, where different sublexical correspondences make different predictions about how the nonword

should be pronounced. In German, these are context-insensitive correspondences (vowels tend to be pronounced as short), context-sensitive correspondences (a vowel followed by one consonant is pronounced as long; a vowel followed by two or more consonants is pronounced as long; Perry, Ziegler, Braun, et al., 2010; Ziegler et al., 2000), and body-rime correspondences (using nonwords for which the body is consistently pronounced in the same way in all real words, either consistent with the context-sensitive correspondences, or not). In English, the context-insensitive correspondence used was that the letter *a* is pronounced as /æ/, the context-sensitive correspondence that an *a* preceded by a *qu* or *w* is pronounced as /ɔ/, and for body-rime analogy we had isolated all bodies where the *a* is consistently pronounced as /oː/, such as *-ald*. Note that these bodies are confounded with the context-sensitive correspondence that an *a* followed by an *l* and another consonant tends to be pronounced as in "bald". This makes it impossible to estimate the independent contribution of bodies for English nonword reading using this paradigm.

In the model, the dependent variable is the percentage of different responses for each item which is obtained from the behavioural data, which is assumed to be a function of both the statistical distribution for each correspondence within the language, and the strength of reliance on each of the correspondences, which is the variable of interest. The optimisation procedure maximises the fit of the model, such that the model finds weightings for each item that are consistent both with the obtained percentages of different vowel pronunciations and the predictions of the statistical distributions. More details about the procedure, items, and analysis are provided in Schmalz et al. (2014; see pp. 142-146; 157-159, 174-176, of Paper 4). The procedure we used here was identical, except that the items were printed out on flash cards and laminated instead of being presented by a computer.

5.3. Results

5.3.1. Overall reading ability

As seen in Table 1, the Australian participants were on average younger than the German participants. This difference was significant, $t(124) = 2.39$, $p < 0.05$. In terms of speeded single word reading skills, Table 1 shows higher average percentiles of the Australian versus the German children. This difference was marginally significant, $t(124) = 1.70$, $p = 0.09$, and indicates that the Australian children were slightly better readers than the German children, given their grade. For their absolute speeded word reading ability as measured by the GEco, the Australian children outperformed the German children, $t(124) = 2.68$, $p < 0.01$.

Thus, even though the Australian children were younger than the German children, they were slightly better readers given their age, and as a result showed better performance on speed-reading a list of relatively high-frequency words (the GEco). We did not attempt to strictly match the children across orthographies on their age or reading ability: this would have required arbitrary decisions, for example, whether to match the children on their age or number of years of reading instruction (which typically starts at age 5 in Australia, and age 7 in Germany); whether to match them on word reading or nonword reading, accuracy, speed, or comprehension, and so on (see also the Thesis Discussion section). The finding that the Australian children perform better on the speeded word reading task than the German children goes against our predictions for the subsequent analyses on the speed of acquisition of lexical and sublexical processing, where we expect the German children to outperform Australian children, due to the transparency of their orthography. However, it should be noted that the differences in age or grade are possible confounding factors: if we find worse performance for Australian than for German children, it might be attributable to those general maturation factors. For

the nonword optimisation task, we do not perform any direct cross-linguistic analyses, therefore the difference in reading ability does not affect the validity of the results from this task.

5.3.2. Lexical and sublexical processing

The performance on the nonwords and irregular words was scored as the number of items read correctly out of 17. The average performance divided up by grade and language is presented in Table 2.

Table 2.

Accuracy of irregular word and nonword reading per grade and language; mean (SD).

	Nonword		Irregular words	
	German	English	German	English
Grade 1	NA	7.78 (5.38)	NA	6.44 (3.78)
Grade 2	11.62 (2.57)	9.16 (3.61)	7.88 (4.28)	9.24 (4.01)
Grade 3	13.26 (2.68)	9.94 (4.37)	12.42 (3.63)	10.88 (3.67)
Grade 4	13.62 (2.31)	10.91 (4.11)	13.33 (2.58)	10.63 (2.16)

The table shows that for nonwords, the performance of the German children exceeded that of the Australian children for all age groups. An ANOVA with grade and language as between-subject factors and by-subject nonword reading accuracy showed a main effect of language, $F(1,126) = 19.01$, $p < 0.001$, $\eta^2 = 0.14$, reflecting better performance of German than Australian children, and a main effect of grade, $F(3,126) = 2.88$, $p < 0.05$, $\eta^2 = 0.07$. The interaction between language and grade was not significant, $p > 0.8$. [20]

[20] The pattern of results was virtually identical regardless of whether we included the Australian Grade 1 or not.

In regards to the irregular word reading performance, Table 2 shows a cross-over, where younger English children outperform the younger German children, but the pattern reverses in Grades 3 and 4. In the ANOVA analysis, using language and grade as the independent variables and by-subject irregular word reading accuracy as the independent variable, we found a main effect of grade, $F(3,126) = 9.64$, $p < 0.001$, $\eta^2 = 0.20$, reflecting an increase in accuracy across grades, and an interaction between grade and language, $F(2,126) = 3.33$, $p < 0.05$, $\eta^2 = 0.05$, but no main effect of language, $F(1,126) = 1.90$, $p > 0.1$, $\eta^2 = 0.02$. This was most likely driven by the cross-over shown in Table 2.

5.3.3. Nonword reading and optimisation

As discussed in Schmalz et al. (2014; i.e., Paper 4), drawing conclusions about cross-linguistic differences based on direct comparisons of the weightings is inappropriate as we are comparing different items containing different types of correspondences. A valid approach, however, is to conduct all analyses within each language, and then contrast the overall patterns of results across languages on a descriptive level. We therefore analyse the German and English data separately.

5.3.3.1. German

Out of the children who performed this task, 19 were in Grade 2, 19 in Grade 3, and 17 in Grade 4. Their nonword pronunciations were scored by the first author (a native German speaker) as being pronounced with a long vowel, a short vowel, or incorrectly. The summary of the children's responses is presented in Table 3. Incorrect responses (i.e., those with lexicalisations, letter substitutions, additions, or deletions), were discarded from the analysis (9.33%).

Table 3.

Proportions of each type of response by grade for German.

Grade	Short vowel response	Long vowel response	Incorrect response	No response
2	61.05%	27.66%	10.88%	0.41%
3	64.62%	26.78%	8.60%	0.00%
4	59.87%	32.16%	7.97%	0.00%

We divided the children based on their grades. The weightings for each of the three Grade groups are shown in Table 4. This table reveals that the reliance on body-rime correspondences increases across grade, while the reliance on context-insensitive correspondences decreases. There is little change in the reliance on context-sensitive correspondences.

Table 4.

Weightings for three sublexical correspondences in German children across grades.

Age group	Context-insensitive correspondences	Context-sensitive correspondences	Body-rime correspondences
Grade 2	0.68	0.25	0.07
Grade 3	0.67	0.21	0.12
Grade 4	0.55	0.23	0.22

The comparison across Grades, as shown in Table 4, cannot distinguish between effects of cognitive maturity and reading experience on reading proficiency *per se*. Therefore, it does not address the question of individual differences in reading ability, and how they relate to the use of each of the three correspondence types. In the next analysis, we calculated the weightings and reading percentile (based on the SLRT standardised word fluency test; Moll & Landerl, 2010) for each individual child to explore the association between different types of correspondences and a child's standardised reading

ability. This controls for differences associated with grade (e.g., age, reading experience), and identifies children as better or poorer readers compared to their peers.

When considering reliance on the three sublexical correspondences, the only significant correlation was between reading percentile and the degree of reliance on body-rime correspondences, where poorer readers relied *more* on body-rime correspondences than better readers, $r(54) = -0.35$, $p < 0.01$.

5.3.3.2. English

The sample of children who completed this task consisted of 8 children in Grade 1, 25 in Grade 2, 14 in Grade 3 and 11 in Grade 4. The English nonword responses were scored by a research assistant with a background in phonology, who was also a native speaker of Australian English. She was unaware of the aim of the experiment. We excluded all non-responses (3%). Responses were scored as correct if the consonant preceding and succeeding the vowel were correct, all others (18%) were excluded from the analysis. We adopted this lenient marking criterion because excluding all items with any errors would have reduced the amount of useable data considerably. We therefore chose to retain all responses that did not disrupt the context of the grapheme-phoneme correspondence of interest (i.e., of the vowel). One participant from Grade 1 had no useable responses for one item type, and was excluded from the analyses. The response types across the participants who were included are summarised in Table 5.

Table 6 presents the by-grade weightings for each of the three types of correspondences. Unlike the German data, this shows no clear trends across grades. Overall, we find little reliance on context-insensitive correspondences, and the strongest influence is from context-sensitive correspondences. This overall pattern is consistent with the responses of English-speaking adults (Schmalz et al., 2014, i.e., Paper 4).

Table 5.

Proportions of each type of response by grade for English

Grade	/æ/ (*cat*) response	/ɔ/ (*was*) response	/oː/ *(bald)* response	Other vowel	Incorrect response	No response
Grade 1	0.38	0.17	0.06	0.11	0.26	0.02
Grade 2	0.43	0.18	0.10	0.10	0.16	0.04
Grade 3	0.38	0.20	0.09	0.08	0.21	0.04
Grade 4	0.43	0.26	0.09	0.09	0.12	0.01

Table 6.

Weightings for three sublexical correspondences in English children across grades.

Age group	Context-insensitive correspondences	Context-sensitive correspondences	Body-rime correspondences
Grade 1	0.04	0.78	0.18
Grade 2	0.09	0.66	0.24
Grade 3	0.25	0.70	0.28
Grade 4	0.001	0.74	0.26

Analogous to the German analyses, we calculated the weightings for each individual child to explore correlations of reliance on different correspondences with the reading percentile, based on the TOWRE speeded word reading test (Australian norms from Marinus et al., 2013). Again, this served to address the possibility that reliance on a particular type of correspondence may co-vary with the status of a child as a good or poor reader given their grade level. In contrast to the German data, reliance on body-rime correspondences was not correlated with the reading percentile, $r(55) = 0.07$, $p > 0.6$. There were, however, significant correlations between reading percentile and reliance on context-insensitive correspondences, $r(55) = -0.40$, $p < 0.01$, with better readers relying less on these simple correspondences, and with context-sensitive rules, $r(55) = 0.31$, $p < 0.05$, with better readers relying more on these complex correspondences.

Unlike the German data in Table 4, the English by-grade weightings from Table 6 do not show any clear developmental trends across the age groups. This is likely due to the relatively good reading skills of the younger Australian readers (see Table 2). We therefore looked at the correlation between overall reading ability, as measured by the TOWRE raw scores, and the weightings. The pattern of correlations was similar when we used TOWRE raw scores instead of percentiles, with a significant negative correlation between reading skill and reliance on context-insensitive correspondences, $r(55) = -0.52$, $p < 0.0001$, a significant positive correlation between reliance on context-sensitive correspondences and reading ability, $r(55) = 0.39$, $p < 0.005$, and no correlation between reliance on body-rime correspondences and reading ability, $r(55) = 0.15$, $p > 0.2$. Thus, the patterns of correlations support the view that better readers rely on more complex rules to a greater extent.

5.3.3.3. Summary of the optimisation results across languages

In comparing the German and the Australian results, two main points emerge: (1) in line with the adult data reported by Schmalz et al. (2014; i.e., Paper 4), German readers rely to the greatest extent on context-insensitive correspondences, while context-sensitive correspondences are the best predictor of vowel responses in English readers; (2) The English data is far less clear than the German data.

5.4. Discussion

The current paper aimed to explore the efficiency of lexical and sublexical processing in reading acquisition across orthographies, as well as the developmental trajectory of the nature of sublexical processing in English and German. Although the English-speaking sample had better speeded word reading performance, the German children outperformed the Australian children on both a nonword reading task and on an irregular word reading task.

Furthermore, we used a nonword reading paradigm in conjunction with a mathematical optimisation procedure to explore the nature of sublexical processing across age and reading ability, in German and English-speaking children. The novel aspect of this procedure is that it allows us to quantify the degree of reliance on three different types of correspondences, and track it as a function of reading experience and ability. Across orthographies, we find convergence with data that has previously been reported with adults (Schmalz et al., 2014, i.e., Paper 4): German children, at all age groups, show the strongest reliance on simple, context-insensitive rules to determine the vowel pronunciation. For English children of all age groups, context-sensitive rules were the strongest predictor. The procedure showed some clear developmental trends for German, while the English data was less clear-cut. This also confirms the findings of Schmalz et al. (2014), that the optimisation procedure is more reliable in German than English.

5.4.1. Efficiency of the lexical and sublexical routes in English and German children

Our first aim was to establish whether lexical and sublexical skills were harder to acquire for English compared to German children. The logic was as follows: due to the complexity and unreliability of the sublexical correspondences in English, and also in line with previous research, we expected that sublexical processing, as measured by nonword reading accuracy, should be less efficient in English than German children. We also expected that the existence of a less efficient sublexical decoding mechanism in English compared to German would impede the build-up of orthographic entries. Note that this would not lead to a cross-linguistic difference for processing high-frequency words: once sound orthographic entries are established (even if it takes longer in English than German), subsequent orthographic access should not be influenced by orthographic depth.

The data was consistent with all of the above predictions. The Australian children, overall, had better speeded word reading ability than German children on the GEco test, which included mostly high-frequency words. The difference in overall speeded word reading ability goes against our predictions, as the Australian children outperform the German children. Therefore, the difference in overall reading ability does not compromise our conclusions: despite good reading ability for familiar words, Australian children still struggle with nonword and irregular word reading relative to the German children.

The novel aspect of this study is the use of irregular words in a cross-linguistic comparison. Although previous studies have shown that development of both word and nonword reading ability lags in shallow compared to deep orthographies, only irregular word reading accuracy can give us information about the ease with which orthographic lexical entries are established. Accurate reading aloud of regular words can be achieved either by the use of the lexical or of the sublexical route, so it is unclear how to interpret cross-linguistic differences in word reading performance, unless only irregular words are used. In our study, the irregular words were matched on frequency across languages. As a result, lower accuracy in English compared to German children indicates that given an equal amount of exposure, German children have a greater chance of establishing orthographic representations.

Although our findings from the irregular word reading task are suggestive, they require follow-up studies. Firstly, the difference in irregular word reading accuracy across languages was not reliable for all age groups: when we included grade as an additional predictor, the main effect of language disappeared, because the children in Grade 2 showed a reverse pattern, where English-speaking children outperformed German-speaking children. As we made no a priori predictions about any cross-overs, and due to

the relatively small sample size per grade, it is difficult to interpret this result. The cross-over is likely driven by the relatively good reading ability of the younger Australian children. Therefore, it would be valuable to replicate this finding with a better-matched sample.

Secondly, the item set was rather small, due to the limited number of irregular words (especially in German) that we judged to be familiar to children. We matched across orthographies on frequency, length, and orthographic N, but there is no guarantee that other variables may not have differed across the item sets, that would have made the English irregular words harder than German irregular words. Thirdly, irregular word errors give us no information about the locus of the deficit: failing to read an irregular word correctly could imply either problems with establishing or storing the representation, or with accessing an existing representation.

For future research, in within-language studies it may be valuable to explore how and why irregular words are more difficult to learn than regular words, as this removes the potential confounds associated with systematic differences in orthographic characteristics across languages. Orthographic learning studies allow for stronger control over the characteristics of the words to-be-learned, and of the consistency of the correspondences that underlie these (Taylor, Plunkett, & Nation, 2011; Wang, Castles, & Nickels, 2012). Once the mechanisms of irregular word learning are understood, these can be more easily brought into the perspective of cross-linguistic research. A broad question that could then be addressed is the extent and mechanism by which overall print-to-speech correspondence reliability affects the process of orthographic learning within the natural environment in which reading acquisition occurs.

5.4.2. Nature of sublexical processing in English and German

The second aim on this study was to examine the developmental trajectory of the reliance on different types of sublexical correspondences across orthographies. These broadly converged with the results of a recent study with adults (Schmalz et al., 2014): we find that, on a cognitive level, context-sensitive rules are the most important predictor for vowel pronunciation in English, and context-insensitive rules are the most important for vowel pronunciation in German. This holds true for all age groups, suggesting that such a cross-linguistic difference emerges early during reading acquisition.

The cross-linguistic differences in the reliance on context-sensitive versus context-insensitive correspondences also fits nicely within a new framework of orthographic depth, which has been proposed by Schmalz et al. (under review, i.e., Paper 6). According to this view, orthographic depth consists of two underlying components: the complexity of the sublexical correspondences, as their unpredictability. Within this framework, the findings become very intuitive: English readers rely to a greater extent on context-sensitive rules than German readers, because the statistical regularities in the orthography that underlie the sublexical route are driven by this type of sublexical correspondences.

Concerning the developmental trajectory across languages, the results were less clear-cut. Specifically, the German children showed a clear pattern, while Australian children did not. This, too, is broadly consistent with the findings of Schmalz et al. (2014), that the optimisation procedure is more reliable in German than English. The English data is more noisy compared to German, and includes a relatively high proportion of "other" vowel responses. There are two possible explanations for this pattern. Firstly, this may be due to the phonological complexity of the English language, where even for a trained phonologist it may be impossible to distinguish, for example, between a shortened

/oː/ response or a lengthened /ɔ/ response. As subtle vowel distinctions are critical to our procedure, this may increase the unreliability of the outcome.

A second, more theoretically interesting possibility is that in addition to the three correspondences that were tested, other sources of information are used to determine vowel length in English. This is in line with the relatively large number of "other" vowel responses - as these remain stable and do not decrease across the age groups (see Table 3), it is unlikely that they simply reflect reading errors on the part of the younger children.

The English results did, however, allow us to draw some conclusions about the developmental trajectory of reliance on different types of units. Although the by-Grade analyses did not show any clear patterns, some significant correlations emerged between the children's reading ability and their weightings: Children who were better readers, both in terms of their raw score and the reading percentile, showed less reliance on context-insensitive correspondences, and stronger reliance on context-sensitive correspondences. This is in line with previous work, which has shown that older children are more reliant on context-sensitive rules than younger children (Treiman et al., 2006). This finding indicates that reliance on these context-sensitive correspondences develops over time, while children become sensitive to the regularities that exist between print and speech on levels that were not explicitly taught.

In German, the reliance on context-insensitive correspondences decreased with age, while the reliance on body-rime correspondences increased. This is likely to reflect that with increasing reading experience, children learn to derive more subtle regularities than those they were taught explicitly, namely that vowel length can be partly predicted by relying on body-rime correspondences. The reliance on context-sensitive correspondences did not change across age, indicating that children by the end of grade

two have already developed sensitivity to the more abstract regularity, that vowel length often depends on the number of consonants that succeed it.

An additional interesting pattern within the German results was the reversal of the relationship for reliance on body-rime correspondences: although it *increased* with age, it *decreased* with standardised reading score. Reliance on body-rime correspondences in German children is particularly prominent for older children who are poor readers given their age. The finding that poor readers rely to a greater extent on body-rime correspondences is consistent with a previous study: Ziegler et al. (2003) used the body-N effect as a marker of reliance on bodies, and found that compared to age-matched controls, both German and English children with dyslexia showed stronger body-N effects. Taken together, these findings suggest that body-rime correspondences are used as a compensatory mechanism in dyslexia. The dominant theory of dyslexia proposes that it is caused by a phonological deficit, or a lack of sensitivity to phonemes (Snowling, 2000). Such a phonological deficit may lead to a relative decrease in the use of grapheme-phoneme correspondences and push for increased reliance on body-rime correspondences, even in a shallow orthography like German. As it takes reading experience to pick up these regularities, this mechanism is not available to younger readers if they do not get explicit instruction about bodies; therefore increased reliance on bodies by dyslexic children compared to their peers emerges only in older children.

5.4.3. Conclusion

In the current paper, we addressed two questions about reading acquisition across orthographies: (1) Does the overall print-to-speech correspondence reliability of an orthography affect the efficiency of lexical as well as sublexical processes? (2) How does the reliance on different types of sublexical clusters develop in English and German? To answer the first question, we found that German children were more accurate than

Australian children at nonword and irregular word reading. This suggests building up orthographic representations, is more difficult in English than in German.

To answer the second question, we found that across all age groups, context-sensitive correspondences were most important in predicting vowel pronunciation of English children, while for German children, context-insensitive correspondences emerged as the stable strongest predictor. This suggests that, from an early age, the complexity of print-to-speech correspondences in English encourages children to rely on these more complex sublexical rules.

Paper 6: Getting to the bottom of orthographic depth

This paper is currently under review at the Psychonomic Bulletin and Review. We would like to thank three anonymous reviewers for their helpful comments so far.

Abstract

Orthographic depth has been studied intensively as one of the sources of cross-linguistic differences in reading, and yet there has been little detailed analysis of what is meant by orthographic depth. Here we propose that orthographic depth is a conglomerate of two separate constructs: the complexity of print-to-speech correspondences and the unpredictability of the derivation of the pronunciations of words on the basis of their orthography. We show that on a linguistic level, these two concepts can be dissociated. Furthermore, we make different predictions about how the two concepts would affect skilled reading and reading acquisition. We argue that refining the definition of orthographic depth opens up new research questions, addressing which can provide insights into the specific mechanisms by which language-level orthographic properties affect cognitive processes underlying reading.

6.1. What is Orthographic Depth?

In the study of reading, it is important to establish to what extent findings from reading in one language can be generalised to another, and what particular experimental results are specific to the particular orthography used in the experiments (Frost, 2012a; Share, 2008). In recent decades, cross-linguistic research in reading has focussed particularly on the concept called *orthographic depth* as a source of cross-linguistic orthographic differences in reading behaviour. Broadly speaking, orthographic depth refers to the reliability of print-to-speech correspondences. English is considered to be a deep orthography, as there are often different pronunciations for the same spelling patterns (e.g., “tough” – “though” – “through” – “bough” –“cough” – "thorough" - “hiccough”; Ziegler et al., 1997). Hence, it has often been contrasted with “shallow” orthographies with more reliable correspondences, such as Serbo-Croatian (Frost et al., 1987; Turvey et al., 1984), German (Frith et al., 1998; Landerl et al., 1997; Wimmer & Goswami, 1994; Ziegler et al., 2001), and many others (see Katz & Frost, 1992; Ziegler & Goswami, 2005 for reviews).

The issue of orthographic depth is relevant for a broad range of issues, including reading development, dyslexia, and models of skilled reading. All aspects of reading are intrinsically linked to the characteristics of the orthography, therefore establishing what orthographic characteristics affect reading processes, and the cognitive mechanisms via which this occurs, is important for practical and theoretical reasons. For example, research on reading acquisition has consistently shown that achieving reading accuracy is a slower process for children learning to read in deep compared to shallow orthographies (e.g., Frith et al., 1998; Landerl, 2000; Seymour et al., 2003; Wimmer & Goswami, 1994). To account for these findings, theories of reading acquisition often consider the role of orthographic depth, and the challenges that it poses for young readers (Goswami,

1999; Liberman, Liberman, Mattingly, & Shankweiler, 1980; Ziegler & Goswami, 2005). The very mechanisms that underlie reading acquisition might be important to different degrees depending on orthographic depth: numerous studies have shown differences in the strength of various predictors of reading ability (Caravolas et al., 2012; Moll et al., 2014; Vaessen et al., 2010; Ziegler et al., 2010). Furthermore, behavioural studies suggest that the symptoms associated with developmental dyslexia differ as a function of orthographic depth (Landerl et al., 2013; Landerl et al., 1997; Wimmer, 1996; but see Ziegler, Perry, Ma-Wyatt, et al., 2003), which has obvious practical implications. These behavioural findings are supplemented by neuroimaging data, which has shown cross-linguistic differences in the brain activation patterns during reading in dyslexic compared to control readers (for a recent review, see Richlan, 2014).

In addition, the concept of orthographic depth touches on issues that are central to debates in the reading literature in general, such as the extent to which reading processes are universal or language-specific (Dehaene, 2009; Frost, 2012a; Share, 2008). Previous research suggests that the cognitive processes underlying skilled reading are dependent on orthographic depth (Frost, 1994; Frost et al., 1987; Schmalz et al., 2014, i.e., Paper 4; Ziegler et al., 2001). Determining which aspects of the reading process are universal and which aspects depend on the characteristics of the orthography has been recently argued to be an essential and inevitable step in creating models of reading (Frost, 2012a, 2012b). More specifically, and relating to the concept of orthographic depth, the majority of reading research is based on English. As has been argued elsewhere, this poses a threat to the generalisability of this research, especially since English is considered to be an outlier on the orthographic depth scale compared to other orthographies (Share, 2008). Although orthographic depth is not the only source of variability across orthographies, it has

probably received the most attention in the past decades. Therefore, understanding what it is and how it affects reading processes is of theoretical importance.

In summary, it is clear that orthographic depth is an important concept, and understanding how it relates to reading is pivotal, as it is a strong source of linguistic variability between alphabetic orthographies. Here, we argue that it is currently unclear what precise mechanisms drive these cross-orthographic differences, both on a linguistic and behavioural level. We propose that a more precise definition of orthographic depth is needed for future research. In particular, answering the question, "what is orthographic depth", involves determining, on a linguistic level, what different aspects underlie this concept, and how these can be quantified. Once a clear definition of orthographic depth is formulated, current theories and models of reading can be used to make specific predictions about how each aspect of orthographic depth might affect skilled reading and reading acquisition.

6.2. Definitions to date

6.2.1. Existing definitions of orthographic depth

As orthographic depth has been explored for decades, a number of definitions have been proposed. Originally, the concept was formulated in terms of a compromise between morphological and phonological transparency (Chomsky & Halle, 1968). In orthographies such as English or Dutch, such compromises are necessary, because the languages are morphologically deep, in that the morphemes can have different pronunciations. Therefore, the orthography needs to either convey the morphology, or the phonology of the word: it cannot convey both. For example, in English, the word "heal" and "health" have the same spelling pattern because they are semantically related, even though they have different pronunciations. Thus, English sacrifices orthographic consistency for morphological consistency. In Dutch, conversely, the words "lezen" (to

read) and “lees” (I read) have different spellings, despite being forms of the same verb. This is because the "z" in "lezen" is pronounced as /z/, whereas consonants in the final position of Dutch words are devoiced; therefore, the pronunciation of the final phoneme of "lees" is /s/, which is represented by the grapheme "s". The vowel doubling in "lees" compared to "lezen" occurs because in Dutch, vowels in open syllables are pronounced as long vowels ("le-zen"), and if a long vowel phoneme occurs within a closed syllable (as in the case of "lees"), the vowel letter needs to be doubled. Thus, the Dutch orthography sacrifices phonological for morphological transparency (Landerl & Reitsma, 2005).

Originally, the term "depth" had two levels, relating either to morphological or phonological transparency. In the context of the reading literature, the concept of phonological transparency has received the most attention (Feldman & Turvey, 1983; Frost, 1994; Frost et al., 1987). Katz and Frost (1992), in a review of the Orthographic Depth Hypothesis (ODH), provide an overview of the origins of the term concept of depth, and its relationship to both morphological and phonological transparency. Their predictions about how depth would affect reading processes, however, focused exclusively on the relationship between orthography and phonology - as we will discuss in detail in a later section.

The relationship between orthography and phonology is considered to vary as a continuum (Frost et al., 1987; Goswami et al., 1998; Seymour et al., 2003; Sprenger-Charolles, Siegel, Jiménez, & Ziegler, 2011). This implies that a given orthography can be classified along an orthographic depth continuum. This, however, is only possible if this concept can be defined in a specific way, that would allow for the development of a linguistic quantification scheme. Arguably, this is currently lacking in the available literature to date.

There is agreement that orthographic depth refers to the reliability of the print-to-speech correspondences, but what exactly differs across orthographies and how this should be quantified is less clear. Frost and Katz (1992) list three different aspects of letter-sound correspondences that could help to flesh out this definition of orthographic depth: "Because shallow orthographies have relatively *simple*, *consistent*, and *complete* connections between letter and phoneme, it is easier for readers to recover more of a printed word's phonology prelexically by assembling it from letter-phoneme correspondences." (pp. 71-72). Similarly, in a more recent paper, Richlan (2014) concurs by describing orthographic depth as "the *complexity*, *consistency*, or *transparency* of grapheme-phoneme correspondences in written alphabetic language" (p. 1). What is now needed are studies concerning how these different concepts work, whether they can be distinguished from each other, and how each might be quantified.

We argue that a more specific definition is needed to create an explicit theoretical framework that accounts for the way in which orthographic depth influences reading. In order to conduct meaningful behavioural cross-linguistic studies, the degree of orthographic depth of the orthographies which are being studied needs to be defined *a priori*, preferably using an objective linguistic quantification method. This is particularly important because orthographies differ from each other in many aspects apart from orthographic depth, such as syllabic complexity, orthographic density, or the proportion of mono- versus polysyllabic words in the language. Unless the concept of orthographic depth is formally defined, it is easy to fall into circular reasoning, where any behavioural differences across orthographies are attributed to orthographic depth post-hoc.

Devising a meaningful quantification method bears further challenges, because the quantification scheme needs to retain a link to theoretical constructs; if it does not, it becomes unclear what the quantification method is actually measuring. Therefore, in

order to get to the bottom of orthographic depth, we need to first understand what constructs underlie orthographic depth and whether these are theoretically important. For a linguistic construct, there needs to be enough variability across orthographies to make across-language studies meaningful. Then, on a behavioural level, we need to be able to show a noticeable effect that is directly associated with the concept of study.

6.2.2. Orthographic Depth in Theories and Models of Reading

Two theories of reading across languages that are primarily concerned with orthographic depth are the Orthographic Depth Hypothesis (ODH; Katz & Frost, 1992) and the Psycholinguistic Grain Size Theory (PGST; Ziegler & Goswami, 2005). Both postulate how orthographic depth would affect reading processes, the ODH with a focus on skilled reading, and the PGST with a focus on reading acquisition, and provide some definition of what is meant by orthographic depth. Katz and Frost (1992) distinguish between three concepts that underlie orthographic depth: they state that in a deep language, the print-to-speech correspondences are *complex, inconsistent,* and *incomplete*. It is unclear, however, precisely how each of these three aspects relates to each other, and whether each of them influences reading in different ways. Katz and Frost (1992) say the following about the specific mechanism that affects reading processes:

> We would like to make two points, each independent of the other. The first states that, because shallow orthographies are optimised for assembling phonology from a word's component letters, phonology is more easily available to the reader prelexically than is the case for a deep orthography. The second states that the easier it is to obtain prelexical phonology, the more likely it will be used for both pronunciation and lexical access. Both statements together suggest that the use of assembled phonology should be more prevalent when reading a shallow than when reading a deep orthography (p. 71, Katz & Frost, 1992).

According to this quote, *complexity* of the print-to-speech correspondences is the key variable driving behavioural differences across orthographies in the ODH framework: assembling the pronunciation is slowed down by the presence of complex correspondences (Rastle & Coltheart, 1998; Rey, Jacobs, Schmidt-Weigand, & Ziegler, 1998), which gives more time for the lexical route to access the lexical information, before the sublexical computation of the pronunciation is complete.

It is unclear how the quote of Katz and Frost (1992) applies to the concepts of consistency and completeness. In English, an example of an *inconsistent* sublexical unit is the letter string “ough”, which can be pronounced in six different ways for monosyllabic words alone. For an inconsistent word, sublexical information is not sufficient to determine the pronunciation – instead, the lexicon must be consulted in order to determine how to pronounce a word containing the inconsistent correspondence (e.g., "though" and "through", which contain nearly identical sublexical information, but have different vowel pronunciations).

The third concept introduced by Katz and Frost (1992) is *incompleteness* of the sublexical correspondences. In English, examples of words with incomplete sublexical information are heterophonic homographs. A heterophonic homograph, such as the word “wind”, has two different pronunciations, each of which is linked to a different meaning. The sublexical information is incomplete, as sentence context is needed to activate both the correct phonology of the word, and the correct semantic representation. In an orthography such as Hebrew, this presents a routine computational problem: here, vowels are mostly not represented in written texts. Many words have identical consonant constellations, and as a result vowel information is needed to tell them apart: for example, the consonant string “DVR” can be pronounced, among other alternatives, as “davar”, meaning “thing” or “dever”, meaning “pestilence” (Frost & Bentin, 1992). Again, it

might not be the case that incompleteness slows down the sublexical procedure, as the sublexical procedure is insufficient to read aloud those items. Instead, semantic and lexical information need to be consulted in order to retrieve the correct pronunciation. If such scenarios of inconsistency and incompleteness occur routinely in the orthography in which a child is learning to read, this may lead to qualitatively different cognitive strategies underlying reading.

In the case of complexity, in contrast to consistency and completeness, it is not *necessary* to rely on lexical information in order to obtain information about a word's pronunciation or semantics. In other words, complexity may lead to a quantitative change in the reading processes, but does not force reliance on lexical-semantic strategies, as inconsistency and incompleteness may do. Thus, the distinction drawn by the ODH between different aspects underlying orthographic depth requires further consideration and empirical work. In particular, it is of interest - and so far, to our knowledge, unexplored - whether these three components would affect reading processes in different ways. This would question the utility of the concept of orthographic depth as a unified construct, and instead support the view that it consists of different sub-components.

The PGST, like the ODH, emphasises the role of complex correspondences in driving cross-linguistic differences in the difficulty of acquiring a given orthography. According to the PGST, children learning to read in a deep orthography attempt to minimise the unreliability of their sublexical correspondences by relying on larger units because these tend to be more predictive of a word's correct pronunciation (at least in English; Peereman & Content, 1998; Treiman et al., 1995). As a result, children learning to read in a deep orthography need to learn a greater number of correspondences: children in a hypothetical perfectly shallow orthography can simply learn the letters and their corresponding sounds, and decode all words with perfect accuracy using only those small

units. According to this view, the necessity to learn many print-to-speech correspondences in deep orthographies slows down the process of reading acquisition, leading to the well-established behavioural pattern where children learning to read in a deep orthography lag behind children learning to read in a shallow orthography based on word and nonword reading tasks (Frith et al., 1998; Landerl, 2000; Seymour et al., 2003; Wimmer & Goswami, 1994).

Looking at both the ODH and the PGST, it then seems that orthographic depth can be described as the existence of *complex* rules that are needed in order to decode new words in a given orthography. Yet, we argue that this is not the whole picture: as Katz and Frost (1992), point out, other properties of print-to-speech correspondences associated with orthographic depth relate to their *inconsistency* and *incompleteness*. Here, we focus on understanding how inconsistency can be defined, as it is more relevant to most alphabetic orthographies.

Generally, consistency relates to the presence of more than one pronunciation for a given letter string. It can be defined either on the level of a grapheme[21] (e.g., "ea" is an inconsistent grapheme, because it can be pronounced as in "bread" or "leak"), or of a body (e.g., "-eak" is an inconsistent body, because it can be pronounced as in "break" or "leak")[22]. The consistency terminology is generally associated with connectionist models (Harm & Seidenberg, 1999; Plaut et al., 1996; Seidenberg & McClelland, 1989); dual-route models (M. Coltheart et al., 2001) are more concerned with the concept of regularity, which is defined as compliance to a set of predetermined grapheme-phoneme correspondence rules.

[21] By “grapheme” we mean a letter or group of letters that is the written representation of a phoneme.

[22] Although consistency can be defined at these two different levels, virtually all the research on effects of consistency on reading (e.g. Jared, 2002) has focused just on body-level consistency.

As we explain below, this distinction is important because the two classes of models make different assumptions about how speech is computed from print. However, both classes of models agree on two points: (1) That there is a non-lexical procedure which uses knowledge of print-to-speech regularities to assemble a word's pronunciation. This is particularly important for reading nonwords (in an experimental setting) and unfamiliar words (in a real-life setting) (2) and that there are some words for which the correct pronunciation cannot be computed using this sublexical routine (e.g., "meringue", "colonel"). In order to adopt a theoretically neutral framework, we use the term *unpredictability* to refer to the degree to which this non-lexical reading route, essential for reading nonwords aloud, correctly translates the words of the orthography from orthography to phonology.

The dissociation between *complexity* and *unpredictability* on a linguistic level is not straightforward in English. This can be illustrated with the example of the minimal word pair: “gist” and “gift”. Arguably, the pronunciation of the word “gist” is transparent, because it can be determined using the context-sensitive correspondence that a “g” followed by an “i” is pronounced as /dʒ/. Alternatively, the pronunciation of the word “gift”, could also be argued to be transparent, if we instead apply the simpler rule that the letter “g” is pronounced as /g/. Therefore, the pronunciation of the word “gist” can be resolved by the use of a complex (context-sensitive) correspondence, but in the English orthography it is also, to some degree, *unpredictable* whether this complex rule will apply or not.

The “gift” – “gist” example shows that in English, the complexity and unpredictability of sublexical correspondences are related and confounded, and indeed it is difficult to dissociate the two. This is not always the case in other orthographies, however. Both Italian and French contain the “g[i]” context-sensitive correspondence. In

both these orthographies there is no unpredictability regarding this rule, as it always applies, meaning there are no words with the pattern “gi” where the “g” would be pronounced as /g/. As we will show later, this is important: an orthography which contains many complex rules which are entirely predictable, is different from an orthography which contains many complex rules but also a great deal of unpredictability. We propose that complexity and unpredictability are two related but linguistically and theoretically dissociable concepts. Thus, we argue that orthographic depth, in the context of European orthographies, is a conglomerate of two separate concepts, namely the complexity of sublexical correspondences and the unpredictability of words' pronunciations given these correspondences. Before we introduce linguistic analyses to provide further support for this idea, we expand further on some theoretical issues about defining these concepts within the framework of existing theories and models of reading.

6.2.3. Defining print-to-speech correspondences

As described above, all computational models of reading include some kind of mechanism that uses knowledge of the statistical regularities between print and speech in a given orthography to assemble a word’s pronunciation. The implementation thereof varies considerably, and is a source of debate between computational modellers (M. Coltheart et al., 1993; M. Coltheart et al., 2001; Perry et al., 2007; Perry, Ziegler, & Zorzi, 2010; Plaut et al., 1996; Seidenberg & McClelland, 1989). The Dual Route Cascaded model of reading (M. Coltheart et al., 2001) contains sublexical rules which are defined as the phoneme which corresponds most frequently to a given grapheme. The rules are position-specific: each rule is either valid for all positions (“t” → /t/ - at least for monosyllabic words), or for the beginning, middle, or end positions (e.g., “y” → /j/ in the beginning of a word, and “y” → /ai/ in the end of a word). These are programmed manually, thus making the DRC a static model of skilled reading. This is the first aspect

which distinguishes the DRC from other models of reading, as others include a learning mechanism that derives the sublexical correspondences from whole-word distributions.

The second major difference between the DRC and triangle models (Harm & Seidenberg, 1999; Plaut et al., 1996; Seidenberg & McClelland, 1989) is that in the DRC, all sublexical correspondences are grapheme-phoneme correspondence (GPC) rules. Examples of GPC rules are “t” → /t/ or “th” → /θ/. GPC rules can also be context-sensitive: for example, "g" is pronounced as /dʒ/ when followed by an "i" ("g[i]" → /dʒ/).

In contrast to the DRC and its GPC rules, triangle models develop sensitivity to units that are larger than graphemes, thereby also showing sensitivity to marker effects that are associated with “large” units, such as bodies. CDP+-type models (Perry et al., 2007; Perry, Ziegler, & Zorzi, 2010) represent a compromise between the DRC and triangle models: although the sublexical route is based on graphemes, it also develops sensitivity to the letters surrounding a given correspondence due to learning in a two-layer associative network.

In the light of our proposal, that complexity and unpredictability are different concepts, the reliance on letter clusters that are larger than graphemes in the triangle models blurs this distinction: Given sufficient training, the orthography-to-phonology mapping process will establish orthography-phonology connections (e.g., Plaut et al., 1996). This will make even a word like "ache" (cf. "cache") predictable (via the whole-word correspondence that "ache" maps onto /æɪk/). This means that in this type of model, there are two ways of resolving ambiguities in the print-speech correspondences in the orthography-to-phonology route: (1) by developing reliance on larger grain-sizes, such as body-rime correspondences, to compute the pronunciation sublexically (e.g., "-alk" --> /o:k/), or (2) if the pronunciation is not predictable based on any sublexical unit, by relying on orthography-to-phonology mappings of the whole word. We consider (2) to be

a qualitatively different strategy from (1), because whole-word information is directly linked to lexical-semantic processes, whereas sublexical units (such as syllables or bodies) are not.

Here, for the sake of simplicity, we adopt a definition of orthographic complexity which is in line with the DRC terminology: we refer to a correspondence between orthography and phonology as complex if either the orthographic element involved consists of more than one letter (e.g., “th” → /θ/), or if the correspondence is context-sensitive (“g[i]” → /dʒ/), or if both are true ("ch[r]" → /k/). This does not mean that our framework is incompatible with other models of reading, as modellers from all perspectives would agree that there is a difference between simple correspondences, where a letter is always pronounced in the same way, and complex correspondences, where a preceding or succeeding letter needs to be taken into account.

Unpredictable words, within the framework of both the DRC and connectionist models, can be defined as words where the sublexical route provides an incorrect pronunciation. In the DRC framework, such words are termed *irregular* words (Andrews, 1982; V. Coltheart & Leahy, 1992; Content, 1991; Parkin, McMullen, & Graystone, 1986; Rastle & Coltheart, 1999; Roberts et al., 2003; Schlapp & Underwood, 1988; Schmalz et al., 2013; Ziegler, Perry, & Coltheart, 2003). Within the connectionist models the concept of consistency is stressed. Although consistency differs from regularity, in the context of the current review it reflects the predictability of a word: a consistent word is defined as one where "its pronunciation agrees with those of similarly spelt words" (Plaut et al., 1996, p. 59). This reflects the mechanisms by which the pronunciation of a word is assembled in a connectionist model: as the sublexical route operates based on statistical regularities which are derived from the print-to-speech correspondences of real words,

unpredictable words in this framework are those which have a different pronunciation to similarly spelled words.

In summary, as a working definition, we refer to complex correspondences as those that are multi-letter ("th" → /θ/) and/or context-sensitive ("g[i]" → /dʒ/), and to unpredictable words as irregular words given the set of GPCs that are implemented in the DRC. Given that the definitions are arguably biased towards the DRC framework, we seek for convergence in alternative approaches for all our findings in the following section.

6.3. Quantifications of orthographic depth

6.3.1. Existing measures of orthographic depth, and their relation to complexity and unpredictability

Having distinguished between complex versus unpredictable correspondences, we can attempt to devise a quantification method for each of these on a *linguistic* level. If these represent two separate concepts underlying orthographic depth, the first step is to demonstrate that they vary independently across orthographies. This will firstly show whether there is enough independent variation of the two concepts to warrant practically meaningful investigation on a behavioural level, and secondly will provide some insights as to how orthographic depth may be quantified. Although large-scale linguistic corpus analyses are outside the scope of the current review, we provide some suggestions which can be expanded on by future work. We discuss and expand on previous quantification methods, and consider their advantages and disadvantages. In terms of demonstrating that there is a dissociation between complexity and unpredictability, we refer to a computational-model-driven approach (Ziegler et al., 2000), and a linguistic-corpus-analysis approach (van den Bosch, Content, Daelemans, & de Gelder, 1994). We also discuss two commonly taken approaches to determine the relative depth of a given

orthography, that are arguably limited by not taking into account the distinction between complexity and unpredictability, namely ranking of orthographies that is not based on an objective linguistic quantification measure (e.g., Frost et al., 1987; Seymour et al., 2003), and the onset entropy measure (Borgwaldt, Hellwig, & de Groot, 2004; Borgwaldt et al., 2005).

Given our DRC-based definitions of complexity and unpredictability, it is intuitive to start by using the existing versions of the DRC across orthographies as an attempt to illustrate cross-linguistic differences given the GPC rules. Specifically, we can simply take the numbers and proportions of complex rules, and the proportion of irregular words in the DRCs of the orthographies in which it has been implemented. The number of complex rules (those which are multi-letter and/or context-sensitive) is a measure of complexity as per our working definition.[23]

This approach of comparing the number and types of GPC rules across orthographies, and the degree to which they are sufficient to read aloud words in a given orthography, has been also taken by Ziegler et al. (2000) when they implemented the DRC in German. They found that both the number of rules - and especially the number of complex rules - and the percentage of irregular words was higher in English than in German. This is in line with the general consensus that German is a shallow orthography,

[23] The working definition ignores two additional source of GPC complexity. The first stems from the position-specificity of rules. As position-specific rules are implemented as separate GPC rules, as opposed to rules which apply to all positions, the presence of position-specific GPC rules inflates the number of rules overall, and therefore this source of complexity is reflected in the total number of rules (see Table 1). A second source of GPC complexity is the presence of phonotactic rules, which are context-dependent rules where the influence of a certain phoneme which precedes or succeeds a letter influences its pronunciation. Such rules are particularly common in some orthographies, such as Russian, but have not been researched to a great extent in other orthographies. For example, neither the current implementation of the German, Dutch, or Italian DRC contain any phonotactic rules, even though these might be more suitable to describe numerous aspects of the print-to-speech conversion. We therefore excluded phonotactic GPC rules in Table 1.

and English is deep. The DRC has also been implemented French (Ziegler, Perry, & Coltheart, 2003), Dutch, and Italian (C. Mulatti, personal communication, 25. May 2014), which allows us to list the numbers, proportions and types of rules in these four DRCs[24]. The results of this analysis are presented in Table 1.

Table 1 shows, as expected, that English is a "deep" orthography, in that it has many rules, and a particularly high percentage of irregular words, while Dutch and German are "shallow", in that they have few rules and a small proportion of irregular words. Interestingly, the DRC approach places the French orthography at one end of the continuum for the number/percentage of complex rules (complexity) - according to which French appears to be even more complex than the English orthography - and at the other end of the continuum for the percentage of irregular words (unpredictability) - where French appears to be even more predictable than German and Dutch. This shows that the distinction between the two concepts is meaningful, as they are not perfectly correlated between orthographies. The Italian DRC shows an even smaller number of rules, and a larger proportion of single-letter rules, compared to both German and Dutch.

Although the DRC approach offers insights into the relative positions of the four orthographies on the two continua, there are three reasons why this approach is limited. Firstly, current versions of the DRC are based on monosyllabic words only. In some languages, the proportion of monosyllabic words is relatively high; for others, polysyllabic words form the majority of all words. This is problematic for across-language comparisons. Furthermore, even in languages where monosyllabic words are frequent, structural properties vary between monosyllabic and polysyllabic words.

[24] Table 1 lists the *number* as well as the *percentages* of each type of GPC rule. From a theoretical standpoint, saying that a language has a large number of rules is different from saying that the rules of a given language are complex. In the set of orthographies that we present here, the number of rules and percentage of complex rules are highly correlated, therefore in practice they cannot be used to dissociate between the two different measures.

Therefore, monosyllabic words are not a perfectly representative sample of anyorthography (for a review, see Protopapas & Vlahou, 2009). Although the DRC

Table 1

Measures of complexity and unpredictability for Dutch, English, French, German, and Italian based on the DRCs, and from van den Bosch et al., (1994) for comparison

Measure	**Dutch**	**English**	**French**	**German**	**Italian**
Total number of rules (DRC)	99	226	340	117	59
Single-letter rules (DRC)	50 (50.5%)	38 (16.9%)	46 (13.5%)	31 (26.5%)	19 (32.2%)
Multi-letter rules (DRC)	38 (38.4%)	161 (71.2%)	218 (64.1%)	55 (47.0%)	8 (13.6%)
Context-sensitive rules (DRC)	11 (11.1%)	27 (11.9%)	76 (22.4%)	31 (26.5%)	32 (54.2%)
% Irregular words	7.0	16.9	5.6	10.5	NA
Parsing accuracy (%)	21.3	24.5	12.9	NA	NA
Generalisation accuracy (%)	81.4	54.3	89.1	NA	NA

Note: For the DRC, the numbers represent the number of rules of each type, and the percentage out of the total number of rules in brackets. Results for parsing accuracy and generalisation accuracy (defined on page 20) are taken from van den Bosch et al. (1994)

approach may still be useful to determine the relative position of each orthography in terms of orthographic complexity and unpredictability, it would be valuable to replicate these findings with an approach which is not limited to monosyllabic words.

Secondly, the cross-linguistic versions of the DRC were implemented independently of each other, without the aim of comparing them directly to each other. For example, the number of words in the DRC's lexicons varies extensively, with 4583 words for Dutch, 8027 words for English, 2245 words for French, and 1448 words for German. This links back to the previous point: the varying number of words in any DRC model is likely to reflect the relative percentage of monosyllabic words in each language. It is unclear to what extent the words in the lexicon are representative of words in each orthography, therefore the DRC results alone should be interpreted with caution.

Thirdly, it is not established that the GPC rules that are implemented in the DRC have full psychological reality. Indeed, there is evidence that other orthographic units are used during reading in English (Glushko, 1979; Treiman et al., 2003), German (Perry, Ziegler, Braun, et al., 2010; Schmalz et al., 2014) and French (Perry, Ziegler, & Zorzi, 2014). Although this does not mean that the DRCs cannot be used as a tool to capture linguistic variability in the complexity and predictability of print-to-speech correspondences by using GPCs and irregular words as a proxy, it is, again, desirable to find converging evidence from a different approach.

Such converging evidence can be found from a computational study of a linguistic corpus of English, Dutch, and French (van den Bosch et al., 1994). Importantly, the corpuses used in this study included polysyllabic words as well as monosyllabic words. In addition, this paper predates the DRCs, and has not been conducted within the framework of the any particular theory or model. The approach of this paper was data-driven, and the authors made no *a priori* predictions about the results.

Van den Bosch et al. (1994) conclude that orthographic depth can be dissociated into two separate measures: the difficulty of parsing letter strings into graphemes on the one hand, and the degree of redundancy in the print-to-speech correspondences on the other hand. The former - which we equate with our concept of complexity - was measured by applying a computationally obtained parsing mechanism to a set of test words. In an orthography with simple correspondences, parsing is easier, because in many cases parsing a word into letters would enable the correct mapping of graphemes to phonemes: an English word with simple correspondences, like "cat", would be parsed into "c", "a" and "t", which can be mapped correctly onto the phonemes of /k/, /æ/, and /t/; a word with complex correspondences, such as "chair" would instead need to be parsed as "ch" and "air", because the constituent letters ("c", "h", etc.) do not map onto the correct phonemes. Since for each of the three orthographies the same amount of training was used, differences in parsing accuracy of untrained test words reflect the difficulty of parsing. Parsing accuracy, overall, was low, indicating that all three orthographies are characterised by high complexity. French showed the lowest level of accuracy, while Dutch and English were at approximately the same level (see Table 1).

For quantifying the degree of redundancy, van den Bosch et al. (1994) report the generalisation performance, or the number of test words pronounced correctly by a computationally obtained set of print-to-speech correspondences in the three orthographies. In order to obtain these correspondences, they first derived all possible print-to-speech correspondences of all sizes (ranging from single letters to whole words). Then they compressed the set of correspondences for each orthography to reduce the redundancy among these rules (e.g., knowing the correspondences "a" → /æ/ and "t" → /t/, as well as "-at" → /æt/, is redundant; knowing the correspondences "a" → /æ/, "l" → /l/, and "m" → /m/, as well as "-alm" → /ɐːm/ is not).

The results showed that both Dutch and French outperformed English, meaning that there are many English words that do not comply with these rules. The generalisation measure is reflective of unpredictability: given the set of correspondences that were defined through the compression algorithm, a large number of words in the English orthography were still unpredictable. The predictability, according to this measure, was higher in Dutch and French than in English. The summary of both the variables is presented in Table 1.

To our knowledge, the quantification scheme of van den Bosch et al. (1994) has not been used to study behavioural differences in the effects of orthographic depth, nor has it been applied to other orthographies. This is an important direction for future research. For our current purposes, it is particularly interesting that the two concepts van den Bosch et al. (1994) suggest as underlying orthographic depth based on their linguistic-computational analysis are consistent with the results of the DRC, and our distinction between complexity and unpredictability.

The case of French is particularly interesting: Both the DRC approach and the analysis of van den Bosch et al. (1994) classified French as a relatively complex orthography (many complex GPC rules, low generalisation performance) - even compared to English. Conversely, both approaches classified French as the most predictable orthography - even compared to Dutch. In previous work on orthographic depth, French has often been described as an intermediate orthography (Goswami et al., 1998; Paulesu et al., 2001; Seymour et al., 2003; Sprenger-Charolles et al., 2011). The French orthography, therefore, shows the importance of distinguishing between the two concepts, as a failure to do so provides a different picture.

This intuitive classification of French as an orthography of intermediate depth has been supported by some of the previous quantification schemes, which did not make the

distinction between complexity and unpredictability as separate constructs underlying orthographic depth. For example, Seymour et al. (2003) classified 13 European orthographies based on their degree of depth. [25] They consulted researchers in participating countries and ranked the orthographies in terms of their depth based on a more intuitive approach. This landed French in an "intermediate" position. It seems, therefore, that this intuitive approach "averages out" potentially theoretically relevant distinctions between separate concepts underlying orthographic depth.

A more objective approach, which has been picked up by cross-linguistic researchers, has been introduced by a measure called *onset entropy* (Borgwaldt et al., 2004, 2005). This quantification scheme reflects the number of different ways in which the initial letter of a word, on average, can be pronounced in a given orthography. Initial letters which consistently map onto the same phoneme involve no ambiguity, so they are assigned a value of 0. The more uncertainty there is in the number of possible pronunciations, the higher the entropy value. Borgwaldt et al. (2005) calculated the entropy values for initial letters across orthographies. The average onset entropy for each orthography was then considered to reflect its relative degree of orthographic depth.

This measure has intuitive appeal, and has been used in behavioural studies of cross-linguistic differences (Landerl et al., 2013; Moll et al., 2014; Vaessen et al., 2010; Ziegler et al., 2010). One of its advantages is the focus on the first letter only. Firstly, this eliminates the bias towards monosyllabic words, that is present in the DRC and other approaches. Secondly, it also increases the comparability across orthographies, because words in all orthographies have initial letters (Borgwaldt et al., 2005; Ziegler et al., 2010).

[25] Seymour et al. (2003) draw a distinction between orthographic depth and complexity, but what is meant by complexity is not what we mean by the same term: they refer to the presence of consonant clusters (i.e., *syllabic* complexity), that do not necessarily map onto the same phoneme (e.g., "str"). This does not relate to orthographic depth, but constitutes a different dimension.

Still, neglecting additional information in a word provides other problems. In English, for example, it is often the vowel pronunciation that is unpredictable, and vowels occur more frequently in the middle of a word (Treiman et al., 1995). In French, print-to-speech irregularities occur mostly in the final consonants, which are often silent (Perry et al., 2014; Ziegler, Jacobs, & Stone, 1996).

We provide two examples that show that although the onset entropy measure is a useful first step in quantifying orthographic depth, it confounds orthographic complexity with unpredictability, meaning that it does not provide the whole picture. According to the onset entropy measure, French (with a value of 0.46) is about half-way between English (0.83) and "shallow" orthographies such as Finnish (0.0) and Hungarian (0.17) (Ziegler et al., 2010). This is in line with the intuitive approach taken by Seymour et al. (2003), but it contradicts both the results of van den Bosch (1994) and those of the DRCs (see Table 1).

Another example is the German orthography, which according to onset entropy is relatively deep: the onset entropy value for German is higher (reflecting higher degree of depth) than that of Dutch, Hungarian, Italian, and even Portuguese, and only slightly lower than French (Borgwaldt et al., 2005). This goes against the intuitive notion that German is close to the shallow end of the orthographic depth continuum (Frith et al., 1998; Goswami et al., 2003; Landerl, 2000; Landerl et al., 1997; Seymour et al., 2003; Wimmer & Goswami, 1994; Wimmer et al., 2000; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003), and that Portuguese and French are generally considered to be of intermediate depth (Goswami et al., 1998; Seymour et al., 2003; Sucena et al., 2009).

This counter-intuitive finding can be explained by the distinction between complexity and unpredictability: the relative complexity of the German orthographic system inflates the onset entropy value, despite German’s relatively high degree of

predictability. For example, German words starting with the letter "s" can have the first phoneme /z/, /ʃ/, or /s/. The pronunciation is, however, mostly predictable: in the onset position, when "s" is succeeded by a vowel, it is pronounced as /z/; when it is succeeded by "p" or "t" or is part of the grapheme "sch", it is pronounced as /ʃ/; and in all other cases it is pronounced as /s/. The two examples of French and German show that onset entropy thus has no way of distinguishing between correspondence complexity and unpredictability, and instead "averages out" the two dimensions, thus making French and German appear to be "intermediate" orthographies despite their relatively high predictability.

The intuitive appeal of onset entropy probably lies partly in its ability to summarise an orthography's "depth" in a single number, which can be correlated with various behavioural outcomes. This has been convincingly done in previous studies, which clearly show that orthographic depth has strong effects on reading and the relative strength of predictors during reading acquisition (Landerl et al., 2013; Moll et al., 2014; Vaessen et al., 2010; Ziegler et al., 2010). These findings are valuable, as they stress that orthographic depth is an important concept to study. In order to understand further how it interacts with various cognitive processes, however, we argue that more experimental work is needed, which has to be based on a linguistic quantification method that distinguishes between different constructs that underlie orthographic depth. Otherwise it becomes unclear whether behavioural differences across languages are attributable to unpredictability, or complexity, or both, or some other cross-linguistic difference.

In summary, we have described two separate approaches that both suggest that orthographic depth is not a single concept, but can be dissociated into complexity and unpredictability of the print-to-speech correspondences. One was introduced two decades ago (van den Bosch et al., 1994), but to our knowledge it has not been extended to other

orthographies or formed the basis of behavioural research. For the purpose of the current paper, the study is valuable because the data-driven computational-linguistic study by van den Bosch et al. (1994) led them to the same conclusions as our theory-based DRC approach. This strengthens the position that on a linguistic level, orthographic depth can be dissociated into two separate constructs.

6.3.2. Limitations and open questions for further research

Our definition of orthographic depth was conceptualised with the aim of being specific, as this is essential for an objective classification measure and for precise predictions about behaviour, based on theories and models of reading. The specificity of our definitions comes with the trade-off that it does not capture all sources of cross-linguistic variability. For example, the definition of unpredictability, when defined at the level of the print-speech correspondences, ignores two further sources of unpredictability that exist in alphabetic orthographies, namely incompleteness and irregularities associated with lexical stress assignment.

As discussed earlier, a previous definition of orthographic depth also included the concept of *incompleteness* (Katz & Frost, 1992). In the context of the Hebrew orthography, incompleteness of correspondences is considered to be a characteristic of sublexical correspondences in deep orthographies (Frost, 1994; Katz & Frost, 1992). The insufficient sublexical information, within our framework, makes the pronunciation of a word unpredictable for the sublexical route, as contextual semantic information is required for accessing a single phonological and semantic entry. This source of unpredictability is not captured in any of the discussed quantification methods. Although it is not particularly relevant for the European orthographies, in the case of Hebrew, it might even be the strongest source of orthographic depth (e.g., Frost, 1994). On a behavioural level, the relationship of incompleteness to complexity or consistency of the

sublexical correspondences is still unclear. Given the need to rely on semantic context to resolve this type of unpredictability, it is possible that the incompleteness of the sublexical correspondences presents a qualitatively different problem compared to complexity and consistency, and if this is the case, placing Hebrew on the same continuum as the European orthographies might not be particularly meaningful. Therefore, future research is needed to relate this source of unpredictability to the unreliable print-speech correspondences that underlie a deep European orthography such as English.

Another source of unpredictability that varies across orthographies, but is not captured by any of the previous quantification schemes, is lexical stress assignment. Some orthographies, such as French, have entirely predictable stress assignment, but others, such as English (Rastle & Coltheart, 2000; Seva et al., 2009), Greek (Protopapas et al., 2006), Russian (Jouravlev & Lupker, 2014), and Italian (Cristina Burani & Lisa S. Arduino, 2004; Colombo, 1992), have some ambiguity when it comes to determining the position of the stressed syllable, and lexical-semantic knowledge needs to be recruited to resolve these conflicts. In Russian, for example, the word "замок" (zamok) has a different meaning depending on whether the first or the second syllable is stressed ("castle" when the first syllable is stressed, and "lock" when the second syllable is stressed; a corresponding example in English is *entrance*). However, it is still, to some extent, unclear via what mechanisms stress irregularity affects reading (Sulpizio, Arduino, Paizi, & Burani, 2013), and how it relates to GPC irregularity. Therefore, this leaves open questions for future research: for example, to what extent can stress assignment be predicted across orthographies, how this differs across orthographies, and what cognitive mechanisms are used to resolve ambiguities underlying stress assignment.

Thus, there are linguistic aspects underlying orthographic depth that are not captured by our definition. A further issue that needs to be taken into account is that there are differences between languages and orthographies that are not at all related to orthographic depth. For example, syllable structure differs between Romance and Germanic languages, where syllables in Romance languages such as Italian are characterised by open syllables and rare cases of consonant clusters, whereas Germanic languages, such as German or English, include many consonant clusters (Seymour et al., 2003). English also differs from languages like German and French because there is little inflection: German and French are characterised by a rich morphological system.

The issue becomes even more complicated when we consider non-alphabetic orthographies: both the languages and the orthographic systems of Chinese or Japanese, for example, are so different to the alphabetic orthographies that we consider here, that classifying and comparing them along the same continuum is not possible. In addition to differences in the nature of the process by which speech is computed from print, they differ in terms of the visual complexity, morphological principles, and even definitions of word boundaries (Chang, Maries, & Perfetti, 2014; Cui et al., 2012; Huang & Hanley, 1995; McBride-Chang et al., 2012).

Therefore, we believe that the most valuable studies in terms of getting to the bottom of orthographic depth would involve the following: (1) Cross-linguistic comparisons, where two orthographies which are similar on as many aspects as possible, but different on the particular issue of interest. Given the difficulty in doing this, a comparison involving only two orthographies should not be taken at face value, and needs to be replicated with other orthographies. (2) Within-language studies can be conducted to isolate the particular aspect of orthographic depth that is proposed to drive cross-linguistic differences. For example, Frost (1994) compared marker effects of

lexical-semantic processing in pointed Hebrew, where the sublexical information is complete, to unpointed Hebrew, where the sublexical information for the same words can be manipulated to be incomplete. In line with the Orthographic Depth Hypothesis this showed stronger lexical-semantic marker effects, in a design that controlled for any cross-linguistic differences that may exist in an across-language design.

6.4. Predictions of the new orthographic depth framework for theories of reading

6.4.1. Some key studies within the new framework

The explicit distinction between complexity and unpredictability as two distinct constructs that make up the concept of orthographic depth is a novel contribution of the current review. Previous research has not been conducted with this distinction in mind, therefore, this work is often subject to more than one interpretation, depending on whether behavioural differences are proposed to arise as a function of complexity, or as a function of unpredictability. We review two previous key studies on orthographic depth and illustrate how different conclusions may be drawn depending on how orthographic depth is defined.

A key finding supporting the Orthographic Depth Hypothesis (ODH) comes from a study of the frequency and lexicality effects, and a semantic priming manipulation (Frost, Katz, & Bentin, 1987). The orthographies explored in this study were Hebrew (deep), English (medium-deep) and Serbo-Croatian (shallow). Indeed, there was an increase in the size of the lexical-semantic effects associated with increasing orthographic depth, suggesting stronger involvement of the lexical-semantic route.

Upon closer inspection, this result could be attributed with equal plausibility to either inconsistency or unpredictability. The inconsistent sublexical information of English and the incomplete sublexical information of Hebrew require the involvement of lexical-semantic processing, which might increase their relative importance. But it is also

possible that the complexity of English slows down the sublexical assembly process, thus giving more time for the lexical-semantic processes to occur, and resulting in larger effects associated with lexical-semantic processing. As English differs from Serbo-Croatian both in terms of complexity and unpredictability, and the source of depth is qualitatively different in English compared to Hebrew, we cannot draw any clear conclusions on this question. In future research, this question could be addressed by comparing complex but predictable orthographies, such as French, to simple and predictable orthographies, such as German or Finnish. If increased lexical-semantic processing is associated with unpredictability, but not complexity, we would expect to find similar lexical-semantic marker effects when we hold predictability constant.

Key evidence for the Psycholinguistic Grain Size Theory (Ziegler & Goswami, 2005) stems from a study comparing the size of the length and body-N effects in English and German (Ziegler et al., 2001). The length effect was stronger in German compared to English, and the body-N effect was weaker in German compared to English, suggesting differences in the nature of the sublexical processing underlying reading in the two orthographies. German and English differ to each other on both complexity and unpredictability, so it is unclear which aspects of these writing systems drive the behavioural differences. For example, an increased body-N effect in English compared to German may reflect a difference in the nature of the sublexical processing (as suggested by Ziegler et al., 2001). If the dominant functional sublexical units of English are bodies, this would mean that the sublexical units are more complex. According to this interpretation, the results of the body-N effect reflect a difference in the complexity of sublexical correspondences. An alternative explanation is that the unpredictability of English encourages a qualitatively different reading strategy compared to German, namely an increased reliance on lexical analogy strategies. In this case, a German reader

might tend towards reading words and nonwords via the sublexical correspondences, whereas an English readers relies to a greater extend on similar lexical entries. Thus, English readers might show a stronger body-N effect compared to German readers, because they are facilitated by the presence of orthographically similar words. Again, studies with orthographies which are matched on complexity but differ in terms of predictability (e.g., English - French) or vice versa (e.g., German - French) might be used by future research to dissociate between effects that are associated with each of these two constructs.

In summary, the next step for future research will be to conduct behavioural studies to establish the extent to which the two dimensions affect reading processes. This opens up a plethora of new research questions about the mechanisms via which the variables underlying orthographic depth independently affect reading, or learning to read. We can show that the distinction between complexity and unpredictability of sublexical correspondences is theoretically meaningful if, based on existing models of reading, there are different predictions about how the two constructs affect cognitive processing. To this end, we provide an overview of specific predictions within existing models of reading about how both complexity and unpredictability, as defined above, might affect both skilled reading processes and reading acquisition.

6.4.2. Predictions for complexity and unpredictability in adults

In skilled adult readers, the ODH proposes that complexity slows down the sublexical assembly process, which gives more time for the lexical route to access the relevant word information. Based on computational models of reading, we would indeed expect that the complexity of print-to-speech correspondences should affect the speed of sublexical assembly. Simulations with the DRC have shown that nonwords which contain multi-letter GPCs ("boace") are processed more slowly than nonwords of equal length,

but containing only simple correspondences ("blusp") (Rastle & Coltheart, 1998). This is postulated to occur because the sublexical route in the DRC operates in a serial fashion. When reading an item containing a multi-letter rule, it activates the first letter of the digraph and its equivalent pronunciation. This pronunciation needs to be inhibited once the second letter starts being processed, because the two letters are then parsed into a two-letter grapheme which has a different pronunciation. Behavioural evidence is consistent with the view that multi-letter graphemes take longer to process than single-letter graphemes (Marinus & de Jong, 2011; Rastle & Coltheart, 1998; Rey et al., 1998; Rey, Ziegler, & Jacobs, 2000). Therefore, there is evidence to suggest that increased *complexity* of the sublexical correspondences slows down the speed of sublexical assembly.

To our knowledge, it has not yet been explicitly shown that the slowing-down of sublexical assembly leads to an increased reliance on the lexical procedure, as stated by Katz and Frost (1992), but this is a question that can be easily addressed by future empirical research. One set of predictions that would follow is that, in words containing complex correspondences, lexical and semantic markers such as frequency or imageability effects should be stronger compared to words containing simple correspondences only.

The concept of unpredictability has been addressed by computational modellers in the form of a debate between regularity and consistency. As explained in an earlier section, this reflects that different mechanisms are used by the sublexical routes of computational models to derive the pronunciation of a letter string. Importantly, all modellers agree that within their framework, some words are unpredictable given the mechanisms that are postulated to underlie sublexical decoding (M. Coltheart, 2012; Seidenberg, 2011), and behavioural studies have shown both consistency and regularity

effects (Andrews, 1982; Hino & Lupker, 2000; Jared, 1997, 2002; Jared et al., 1990; Metsala et al., 1998; Parkin et al., 1986; Rastle & Coltheart, 1999; Waters & Seidenberg, 1985; Waters, Seidenberg, & Bruck, 1984), which supports this theoretical proposal. Thus, although different types of models make different predictions about the specific mechanisms via which unpredictability affects reading, all models predict that unpredictability will influence skilled word reading.

In summary, the currently available computational models of skilled reading make predictions about both complexity and unpredictability. In the DRC, the presence of complex rules slows down reading due to a "whammy" effect, where the pronunciation of the single-letter correspondence needs to be cancelled when the context changes the letter's pronunciation. Unpredictability effects, in contrast, occur due to competition between the lexical and sublexical routes. Resolving this conflict *requires* the involvement of lexical-semantic processing. As outlined above, this allows us to make several testable predictions, which can be explored by future research using either within- or across-language designs. This will contribute to the understanding of the precise cognitive mechanisms which drive the cross-orthographic differences that have been previously attributed to the broad concept of orthographic depth.

6.4.3. Theories of reading acquisition and orthographic depth

There are fewer specified models of reading acquisition than models of adult reading. In the case of exploring the effect of orthographic depth on reading acquisition and making specific predictions, computational models could be particularly useful. Connectionist-type models, for example, use a learning algorithm that extracts the regularities in the correspondences between print and speech. Thus, they are faced with a similar problem as a child learning to read (Hutzler, Ziegler, Perry, Wimmer, & Zorzi, 2004). If, for the sake of simplicity, we focus purely on the acquisition of sublexical skills, we can make clear

predictions about complexity and unpredictability. As stated by the Psycholinguistic Grain Size Theory (PGST), the degree of complexity of the sublexical correspondences should make it more difficult to learn these (Ziegler & Goswami, 2005). Applying this to reading acquisition in children, this means that becoming proficient at using sublexical decoding should take longer in an orthography with complex correspondences than in an orthography with simple correspondences.

In terms of learning the sublexical correspondences, we can also make some clear predictions about unpredictability that should be testable both with a connectionist-type model and in children learning to read. If we were to pick two orthographies that are comparable in terms of complexity, but different in terms of predictability, we would expect that learning these correspondences would take the same amount of time. However, after the correspondences are learnt, we would expect that the accuracy in applying these correspondences to new words would be higher in the more predictable orthography.

Both behavioural and computational data from English and German provide some support for this claim. For example, behavioural data has shown that nonword reading accuracy is higher for German than English children (Frith et al., 1998; Landerl, 2000; Ziegler, Perry, Ma-Wyatt, et al., 2003). This holds true even when a lenient marking criterion is used for English, whereby any plausible pronunciation of the nonword is scored as correct. Similar data has also been obtained from comparisons of English with other shallow orthographies, most notably from a large-scale study which included children from 13 European countries (Seymour et al., 2003). A computational study has compared the performance of a sublexical learning algorithm in German and English (Hutzler et al., 2004). In these simulations, the model's nonword reading accuracy of German exceeded that of English, even after a large number of training cycles when the

models had reached a plateau.[26] A limitation of this comparison, and the existing behavioural studies, is that the orthographies that are compared to English differ both in terms of complexity and unpredictability. Therefore, although the existing data is suggestive, we cannot unequivocally attribute the differences in nonword reading accuracy to unpredictability. Future research along these lines would benefit from choosing orthographies which differ in complexity, but are comparable in terms of predictability, such as French and German, to establish the source of the developmental and computational patterns.

Within the concept of unpredictability, we can make predictions about cross-linguistic differences in reading accuracy, but it is also important to bear in mind that the various sub-skills underlying reading do not develop in isolation. In particular, there is a bidirectional relationship between the acquisition of the lexical and the sublexical route (Share, 1995; Ziegler et al., 2014): lexical entries are predominantly established by a self-teaching mechanism which uses knowledge of the sublexical correspondences to decode unfamiliar words, but lexical entries are also used to refine the knowledge of sublexical correspondences. In terms of complexity, we therefore expect that not only the acquisition of the sublexical route, but also that of the lexical route will be delayed in a complex compared to a simpler orthography. The cross-linguistic differences should, in this case, be *quantitative*. In the case of unpredictability, there could be some *qualitative* differences in the mechanisms that are used for self-teaching. Recent within-language studies of orthographic learning provide some support for this notion (Taylor et al., 2011; Wang et al., 2012). In an orthographic learning study, participants are asked to learn new

[26] A simulation study like that by Hutzler et al. (2004) could also be used to test the claims we make about the effect of complexity on learning to read. However, as Hutlzer et al. used a lenient marking criterion to evaluate the models' nonword reading performance, it is hard to make direct comparisons based on their reported data about the speed of acquisition of the sublexical correspondences.

words. These can be assigned either a predictable or an unpredictable pronunciation. Both studies found that when the pronunciation was unpredictable, semantic context facilitated learning. This was not the case for predictable words, where phonological decoding appeared sufficient for orthographic learning. These findings raise questions about cross-linguistic differences in learning to read as a function of unpredictability of sublexical correspondences. It is possible that children learning to read in a relatively unpredictable orthography routinely rely to a greater extent on contextual cues compared to children learning to read in a relatively predictable orthography. This would result in a qualitative shift in the types of cognitive strategies that are used across orthographies differing in predictability to establish orthographic representations. [27]

In summary, given the theories of reading acquisition, we can assume that the two concepts of complexity and unpredictability should affect cognitive processes during learning to read in different ways. Looking purely at the development of sublexical decoding skills, we expect that complexity would slow down the speed of reading acquisition, whereas unpredictability would reduce decoding accuracy, even after all the correspondences have been learned. Furthermore, if children are routinely faced with unpredictable words, it is possible that they need to develop compensatory strategies to achieve high reading accuracy and comprehension. Such a compensatory strategy might be to rely on semantics to a greater extent than do children learning to read in predictable orthographies.

Future research may benefit in particular from studies in naturalistic reading settings, for example by using eye-tracking to monitor the online processes during silent reading of sentences or texts. Such experiments could serve firstly to test the predictions

[27] Wang et al. were working within a DRC framework and manipulated GPC *regularity*. Taylor et al. were working within a more triangle-like framework, and manipulated *body consistency*. Their results and conclusions were strikingly similar.

that we have made based mainly on models of single-word reading, and secondly to establish the extent to which cross-linguistic differences are observable independently of the reading task. Given that silent reading, in contrast to reading aloud, may emphasise lexical-semantic processing, such paradigms might be particularly useful for testing our predictions regarding the involvement of lexical-semantic processing as a function of predictability.

6.5. Conclusions

Behavioural studies of orthographic depth have been conducted for decades, and have shown that it affects the cognitive processing underlying skilled adult reading (Frost et al., 1987; Schmalz et al., 2014; Ziegler et al., 2001), the rate of reading acquisition (Frith et al., 1998; Landerl, 2000; Seymour et al., 2003; Wimmer & Goswami, 1994), the prevalence and symptoms of developmental dyslexia (Landerl et al., 1997; Paulesu et al., 2001; Wimmer, 1993, 1996), brain activation (Paulesu et al., 2000; Richlan, 2014), and the strength of cognitive predictors of reading ability (Caravolas et al., 2012; Caravolas et al., 2013; Landerl et al., 2013; Moll et al., 2014; Vaessen et al., 2010; Ziegler et al., 2010). Clearly, orthographic depth is an important and relevant factor, both for practical and theoretical reasons. We can be confident in concluding that orthographic depth affects reading, but in order to learn more about *why* and *how* this happens a more precise definition of orthographic depth is required. Such a definition is needed to (1) devise a quantification method of linguistic characteristics of the orthographies that is theoretically meaningful, and (2) to use this quantification method for future cross-linguistic research to isolate specific cognitive mechanisms that are affected by the linguistic constructs.

We propose that orthographic depth is a conglomerate of two separate concepts, namely the degree of complexity and predictability of print-to-speech correspondences in a given orthography. We have shown that, on a linguistic level, the two concepts can be

dissociated. Furthermore, given the currently available models and theories of reading, we also expect that each of the two concepts would influence skilled reading and reading acquisition in different ways. Thus, we argue that there are many unanswered questions in the area of cross-linguistic research relating to orthographic depth. These will be able to be pursued more effectively in the context of a systematic framework for orthographic depth.

General Discussion and Conclusion

General Discussion and Conclusions

In the current thesis, we report on five empirical studies and one theoretical study. We aimed to determine how and why the complexity and reliability of the correspondences that exist on a linguistic level (associated with varying degree of orthographic depth) affect the print-to-speech correspondences that form the sublexical route of readers. In the previous literature, two theories of reading across languages varying in orthographic depth have been proposed, namely the Orthographic Depth Hypothesis (Katz & Frost, 1992) and the Psycholinguistic Grain Size Theory (Ziegler & Goswami, 2005). We discuss how the results presented in the thesis relate to various aspects of these two theories in the first section of the discussion.

In the experiments conducted for this thesis, we focussed predominantly on sublexical processing, as previous work has reported that the reliance on different types of sublexical correspondences differs across orthographies varying in orthographic depth (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). We discuss, in turn, the findings of this set of experiments, and the implications for theories of reading development and models of skilled reading. Finally, we provide an overview of the methodological issues relating specifically to cross-linguistic research that we encountered throughout the thesis.

Theoretical implications: Cross-linguistic theories of reading

In the introduction, we described two major theories of reading across languages: the Orthographic Depth Hypothesis (Katz & Frost, 1992) and the Psycholinguistic Grain Size Theory (Ziegler & Goswami, 2005). Some of our papers address issues that are central to the assumptions of these theories, therefore we will consider the implications of our results for each theory in turn.

The Orthographic Depth Hypothesis

According to the Orthographic Depth Hypothesis, depth can be seen as a continuum (Frost et al., 1987). The degree of orthographic depth is proposed to cause a cross-linguistic difference in the cognitive processes underlying reading: the deeper the orthography, the greater the ratio of lexical-to-sublexical processing. Although this has never been stated by the authors of the Orthographic Depth Hypothesis, subsequent research on reading across languages may have created a misleading impression: namely, that all orthographies can be placed on an orthographic depth continuum, and that orthographic depth is the major source of all cross-linguistic differences in reading.

As we argued in Paper 6, it would be impossible to propose a definition of orthographic depth that would incorporate all orthographies: if a definition were broad enough to capture the wealth of cross-linguistic differences that exist in the world's orthographies, it would cease to be useful (Plaut, 2012). Using a more stringent definition of orthographic depth, it quickly becomes clear that the differences in the world's orthographies cannot be reduced to a single dimension: the nature of depth is different in German compared to French (where the difference is mainly in the complexity of the sublexical correspondences), in French compared to English (where the difference is mainly in the predictability of the sublexical correspondences), in English compared to unpointed Hebrew (where English is deep due to complex and unreliable sublexical correspondences, whereas Hebrew is deep because of missing sublexical vowel information), and altogether different in Chinese (where different principles govern print-to-speech translation compared to alphabetic orthographies).

More specifically and in relation to the English-German comparison, the Orthographic Depth Hypothesis proposes that in English, which has a deep orthography, the lexical route should be relatively more important than in German, which has a shallow

orthography. We attempted to test this hypothesis with a cross-linguistic comparison of the regularity effect in Paper 1. As we found that the nature of the irregularities was not comparable across German and English, however, we could not conduct any direct comparisons of the effect across languages. In Paper 2, although our primary aim was not to test any assumptions of the Orthographic Depth Hypothesis, it is relevant that we found a somewhat stable interaction between language and frequency in the reading aloud task: the frequency effect was stronger for English compared to German readers. As the frequency effect is a marker of lexical processing, this finding provides tentative evidence for the claim that lexical processing has a stronger role in English than in German.

The Psycholinguistic Grain Size Theory

The Psycholinguistic Grain Size Theory proposes a qualitative difference in the way that print-to-speech computation occurs in deep versus shallow orthographies (Ziegler et al., 2001). The theory is proposed to apply to all orthographies and to reading in both alphabetic and non-alphabetic scripts: for example, Ziegler and Goswami (2005) consider how the Psycholinguistic Grain Size Theory might apply to orthographies as diverse as Japanese, Korean, and Chinese.

In the case of the German-English comparison, the Psycholinguistic Grain Size Theory suggests that the difference in the nature of sublexical processing lies in the degree to which large or small units are recruited during reading. The theory emphasises the role of body units (Goswami, 1993, 1999, 2002; Goswami et al., 2003; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). The main evidence for cross-linguistic differences in reliance on bodies comes from two lines of research: firstly, two studies on the body-N effect by Ziegler et al. (2001; 2003), and secondly, two studies using nonwords with existent or non-existent bodies with children in English, French, Spanish, and Greek (Goswami et al., 1998; Goswami et al., 1997).

As we did not find any evidence for differential reliance on bodies in English versus German (Papers 1 - 3), we provide a brief overview of how our results may be consolidated with this previous literature. In the case of the Ziegler et al. (2001; 2003) studies, we have discussed several possibilities for the discrepant results throughout the thesis (Papers 1 - 3): there are both item- and participant-level confounds that could provide an alternative explanation for differential reliance on bodies in English and German. Item-level confounds include orthographic density, which leads to systematic differences in orthographic N and body-N across languages unless these are specifically controlled for. Participant-level confounds include methods of instruction (phonics versus whole-word), and overall reading ability.

In the case of the studies of Goswami et al. (1997; 1998), a closer look at their results also provides an alternative explanation for the apparent greater reliance on bodies in deep compared to shallow orthographies. In all their analyses, there were significant effects of language, with children from deeper orthographies reading both less accurately and more slowly than children from shallower orthographies. This may have led to an over-additivity effect, where the absolute difference between the two nonword conditions may be inflated due to a larger overall percentage of errors and longer overall latencies (Faust, Balota, Spieler, & Ferraro, 1999). Although this does not exclude the possibility that reliance on bodies differs across orthographies as a function of orthographic depth, the possibility remains that once these overall differences in reading ability across languages are controlled for, the cross-linguistic differences in the reliance on bodies would disappear.

To further explore the reliance on bodies in English and German readers, we conducted eight body-N experiments with adults (Paper 2), and one experiment with children (Paper 3). In our experiments, English readers did not show a stronger body-N

effect than German children, when item- and participant-level characteristics were controlled for. In Paper 2 with adults, we found no evidence for a body-N effect at all (except for the lexical decision nonwords condition).

Moreover, we further explored the reliance on large units by English and German readers with an alternative marker effect, namely the consistency effect, to measure reliance on body-N (Paper 1). The rationale here is that if reliance on bodies is stronger in English than German, the inconsistency of their pronunciation should affect reading performance to a greater extent in English compared to German. We found a consistency effect, but no language interaction. Together, the findings from Papers 1 - 3 suggest that there are no differences in the reliance on bodies in English compared to German, which calls for a re-examination of the critical claim underlying the Psycholinguistic Grain Size Theory.

While we found no cross-linguistic differences in the reliance on body-rime correspondences, Papers 4 and 5 showed another way in which sublexical processes may differ in English and German. Here, we used a mathematical optimisation method to infer what sublexical correspondences were used by the participants to pronounce a set of nonwords. This way, we could quantify the reliance on context-insensitive versus context-sensitive print-to-speech correspondences in both orthographies.

Both adults (Paper 4) and children (Paper 5) showed cross-linguistic differences in the weighting patterns, where context-sensitive correspondences were the strongest predictor of vowel pronunciations in English, whereas context-insensitive correspondences were the strongest predictor of vowel pronunciations in German.

Overall, our findings are consistent with the Psycholinguistic Grain Size Theory if we consider context-sensitive correspondences to be a type of large sublexical unit. Since we found no cross-linguistic differences in the reliance on body-rime correspondences

(Papers 1 - 3), we propose that the major difference in the nature of sublexical processing between English and German is that in English, it is more important to learn about context- and position-specific regularities when it comes to applying print-to-speech correspondences (Papers 4 - 5).

Due to our focus on context-sensitive correspondences, and in contrast to the Psycholinguistic Grain Size Theory, our proposal cannot be transferred directly to all orthographies. For example, the system underlying the print-to-speech conversion in the Chinese orthography is complex, but not in the sense that it is characterised by context-sensitive correspondences (i.e., an orthographic unit's pronunciation is dependent on a preceding or succeeding sublexical unit). However, if we attempt to generalise the results from Papers 4 and 5 to other orthographies, we can assume that they capture a universal and intuitive principle: namely, that the correspondences that form the sublexical route of a reader in a particular orthography reflect the regularities which can be used to successfully derive the pronunciation of a word in that orthography.

Summary

In terms of the direct predictions of the two cross-linguistic theories for reading in English and German, we hypothesised the following: According to the Orthographic Depth Hypothesis, lexical-semantic marker effects should be stronger for English than German. According to the Psycholinguistic Grain Size Theory, the nature of the sublexical processes underlying reading in English and German should be different, because English readers should rely to a greater extent on large units. We found incidental evidence for the first hypothesis (Paper 2), and evidence for the second hypothesis (Papers 4 & 5), but only when we define "large" units as context-sensitive correspondences. We found no support for the previously reported findings that reliance on bodies differs across orthographies (Papers 1 - 3).

In the light of the definition of orthographic depth that we propose in Paper 6, we would like to interpret these results within a speculative framework, which could form the basis of future empirical study: tentatively, we propose that increased reliance on lexical-semantic processing, as proposed by the Orthographic Depth Hypothesis, is mainly a result of the *unpredictability* of English compared to German. The increased reliance on context-sensitive rules in English compared to German, which we found in Papers 4 and 5, is likely to reflect the *complexity* underlying the regularities between print and speech.

Reliance on body-rime correspondences

The experimental work in Papers 1 - 5 unanimously suggests that readers of both German and English rely on body-rime correspondences. This is true of both adults (Papers 1, 2, & 4) and children (Papers 3 & 5). Previous research has consistently shown that bodies have a psychological reality for readers of English (e.g., Brown & Deavers, 1999; Glushko, 1979; Goswami et al., 1998; Goswami et al., 1997; Treiman et al., 1990).

To date, the evidence for reliance on bodies in German has been sparse, and yielded equivocal results. Ziegler et al. (2003) showed a reliable body-N effect in German children, but the body-N effect for German adults was marginally significant by subjects and not significant by items (Ziegler et al., 2001). Taken together with the results of the thesis, it seems that both adults and children rely on bodies, but the body-N effect is not a sensitive marker for reliance on bodies in adults.

Thus, bodies are important for the reading process, and this is true for both German and English. This finding raises some questions. Firstly, it remains unclear how bodies are processed (for example, lexically or sublexically; see Papers 2 & 3). Secondly, further consideration is required as to why reliance on bodies develops in the first place. We discuss the "how" and "why" questions, in turn, below.

How are bodies processed?

Are bodies lexical, or sublexical, or both?

In Papers 2 and 3, we attempted to identify the locus of the body-N effect, and found somewhat conflicting results. Bodies might be either represented in a sublexical route as body-rime correspondences, or in the lexical route as a hierarchical structure in an interactive activation network (a detailed description of how different models would show sensitivity to bodies is provided in Paper 2, pp. 53-56). We suggested that we can distinguish between these two possibilities by comparing the body-N effect for words versus nonwords, in lexical decision versus in reading aloud, and by its interactions with frequency.

In Paper 2 with adults, we found a body-N effect in only one condition, namely for lexical decision of nonwords, where it was inhibitory. This suggests a lexical locus, as it would appear that nonwords with a high body-N activate many lexical entries, making the nonword harder to reject due to the ensuing lexical activation. In Paper 3, using a reading aloud paradigm with children, we found a body-N effect for nonwords, but not for words, for German monolingual children, but no body-N by lexicality interaction for bilingual children. This suggests a sublexical locus: if bodies are represented as body-rime correspondences we would expect that high body-N would facilitate reading aloud for nonwords, but that for words, lexical processing would override this sublexical effect. Among bilingual participants, there was variability in the degree to which they spoke German or English at home, and we expected this to also lead to lower average familiarity with the spoken word forms of the words that were used in the experiment. Words with low familiarity require greater involvement of the sublexical route, which is consistent with the finding that the bilingual children showed no interaction between lexicality and body-N.

These results contradict each other: one study showed that the locus of the body-N effect was lexical, and the other showed that it was sublexical. One way to resolve this inconsistency would be to refer to a model that does not make a distinction between a lexical and sublexical route, such as Marcus Taft's interactive activation model (Taft, 1991). Here, there is interactive activation between similar words, which are represented hierarchically in a metaphorical lexical space (Forster & Taft, 1994), but also between sublexical units, such as letters, graphemes, and bodies (Taft, 1991).

Finding both a lexical and a sublexical locus could also be explained within a dual route model (M. Coltheart et al., 2001; Perry et al., 2007) - although explaining the pattern of results within this framework requires some post-hoc assumptions. A dual-route model could simulate our results if both the lexical and the sublexical route were modified to become sensitive to body units.

This might seem like over-fitting the model to the explanation, however, it might be plausible if we consider that during reading acquisition, the development of the lexical route is closely linked to the functioning of the sublexical route (Share, 1995; Ziegler et al., 2014). In fact, this might explain why we find evidence for a sublexical locus in children, and evidence for a lexical locus in adults: the way in which words are acquired during the self-teaching stage might influence the structure of the mental lexicon in adulthood.

Parallel processing of different units in the sublexical route

A theoretically important point of Paper 4 is that participants rely on different types of units on different occasions. This suggests that the sublexical route processes various letter parsing options in parallel. Such a parallel mechanism poses a computational problem: firstly, the system needs to make decisions about all plausible parsing options, and secondly, it must resolve a conflict between different output options

if the sublexical units make contrasting predictions about a pronunciation (e.g., in the case of a nonword like *dalm*, the system would activate, in parallel, the options *d-a-l-m* → "/dælm/", *d-al-m* → "/doːm/" and *d-alm* → "/dɐːm/").

Although verbal models have proposed the parallel use of various sublexical units for over a century (e.g., Brown & Deavers, 1999; Huey, 1908; LaBerge & Samuels, 1974; Prinzmetal et al., 1986; Taft, 1991; Ziegler & Goswami, 2005), it is less clear how these conflicts might be resolved in practice. It is essential to consider how a computational model might be altered to be able to more closely mirror participants' behaviour, as this would generate hypotheses as to a specific cognitive mechanism.

The classical implementation of the dual-route framework, the DRC, operates on grapheme-phoneme correspondence rules, and has been criticised by previous research for not showing any reliance on body-rime correspondences (see M. Coltheart, 2012; Perry et al., 2007, for reviews). The idea of reliance on sublexical units that are larger than graphemes is not intrinsically incompatible with a DRC-like model (M. Coltheart et al., 1993; Patterson & Morton, 1985), but it is not straightforward to determine the changes that need to be made to increase the DRC's reliance on bodies. If we adopt the view that reliance on bodies reflects (at least in part) lexical processes, such changes could involve modifying the interactive activation network of the lexical route (Forster & Taft, 1994; Jared, 2002).

If we instead assume that bodies are processed (at least in part) by the sublexical route, then modifying the sublexical route in such a way that it shows simultaneous sensitivity to different types of units will provide a serious challenge. Inserting context-sensitive rules, or even body-rime correspondences into the sublexical route is easily done, but it is not sufficient to simulate the pattern that we report in Paper 4. If "larger" sublexical correspondences are inserted into the sublexical route of the DRC, it will

always rely on the largest possible correspondence, as it is programmed to parse a letter string into the orthographic unit with the largest amount of letters (e.g., when it is presented with the letter string *ough*, it will prefer to parse it as a single unit over the alternative parsings of *o-u-g-h*, *ou-g-h*, etc.). This means that it cannot simulate the finding that the same participant may rely on one type of unit, for example a context-insensitive grapheme, on one occasion, and on another, such as a body, on a different occasion.

In Paper 4, we also simulated the nonword pronunciations with the CDP+ and CDP++ models. Given the two-layer associative network that these models incorporate, one might expect that the models would be able to simulate the behavioural results: the two-layer associate network allows the model to develop sensitivity to the context of each grapheme-phoneme correspondence. However, although the performance of the CDP+ and CDP++ approached the participants' responses to a somewhat greater extent than the DRC did (both the available version of the DRC, and one which included additional context-sensitive rules), it tended to underestimate the extent to which participants relied on context-insensitive rules.

As the quantification scheme we introduced in Paper 4 is just that - a quantification scheme - it does not provide any insights into the mechanisms that may underlie this behaviour. This important question remains open for future research: a thorough computational and behavioural investigation of how participants decide what types of units to rely on in what circumstances would be a valuable contribution to the field.

Why does reliance on body-rime correspondences develop?

Throughout the thesis, we have found evidence for reliance on body-rime correspondences in both English and German, but no cross-linguistic differences that

could be attributable to orthographic depth *per se*, as claimed by the Psycholinguistic Grain Size Theory (Ziegler & Goswami, 2005). This raises the question of the pressures that encourage readers of German to rely on body-rime correspondences. None of our studies were designed to address the "why" question, but previous research has provided insights into the reasons why reliance on larger units may develop. This question is important in interpreting our findings of reliance on bodies in both English and German children and adults, and particularly the evidence that reliance on bodies does not differ in English compared to German (Papers 1 - 3). The literature on the role of bodies in reading has previously rested on the assumption that there are cross-linguistic differences associated with orthographic depth (Goswami et al., 1998; Goswami et al., 1997; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). As we discuss in the section below, our data is not consistent with this explanation. We subsequently consider alternative views, which relate to the role of rime versus phoneme awareness in reading acquisition and the relationship between reliance on larger sublexical units and reading speed.

Reducing unpredictability

Evidence for the reliance on bodies stems predominantly from English-speaking countries (e.g,. Forster & Taft, 1994; Goswami, 1991; Treiman et al., 1990; Treiman et al., 1995). One explanation for the importance of bodies in reading is that in English, bodies are more predictive of vowel pronunciation than graphemes (Treiman et al., 1995). If this is the case, one would expect cross-linguistic differences in the reliance on bodies, depending on the degree to which bodies are more predictive of words' pronunciations in that particular orthography than other sublexical units (Goswami, 1999, 2002; Ziegler & Goswami, 2005). Our finding, that German and English rely on bodies to the same extent, has implications for this explanation.

Reliance on body-rime correspondences in German, by itself, does not exclude the possibility that body units are important because they reduce the ambiguity of a word's pronunciation. The German orthography is not perfectly shallow, as there is often ambiguity in vowel length pronunciation (Perry, Ziegler, Braun, et al., 2010; Ziegler et al., 2000). If, in German, bodies are more predictive of vowel length pronunciation than graphemes, this might push for the reliance on body units.

The possibility that in German, body-rime correspondences may be more predictive of vowel length pronunciation than grapheme-phoneme correspondences, can be tested by a corpus analysis. For the reliability of grapheme-phoneme correspondences, one can take the percentage of irregular words from the DRC, as this shows the percentage of words for which the pronunciation is not predictable, given the correspondences that are implemented in the sublexical route. The German DRC of Ziegler et al. (2000) shows that 90.4% of German monosyllabic words can be decoded correctly via the grapheme-phoneme correspondences. For estimating the reliability of body-rime correspondences, one can calculate the proportion of words where the pronunciation is predictable by body-rime correspondences. An unpublished corpus analysis has been cited as showing that 84% of all German monosyllabic words have consistent bodies (J. Ziegler, unpublished data; as cited by Aro & Wimmer, 2003).

These values are to be interpreted with caution. First, it is unclear to what extent the grapheme-phoneme correspondence rules of the DRC have psychological reality (Perry, Ziegler, Braun, et al., 2010; Schmalz et al., 2014, i.e., Paper 4). Second, the presence of words with inconsistent bodies does not necessarily mean that they are unpredictable: if the majority of inconsistent words have a high friends-to-enemies ratio, their pronunciation can nevertheless be predicted by body-rime correspondences. A more thorough corpus analysis is beyond the scope of the thesis, but future research would

benefit from a cross-linguistic examination of corpuses, to compare the extent to which taking body-rime correspondences decreases pronunciation unpredictability in comparison to grapheme-phoneme correspondences (for an example of a French-English comparison, see Peereman & Content, 1998).

Until more reliable corpus analyses are available, we can make no statements about the absolute gains in predictability when body-rime correspondences are taken into account in German. However, we can estimate the relative degree of grapheme versus body consistency in comparison to English. As noted above, we would predict that the relative degree to which there are gains in predictability should, according to the psycholinguistic grain size theory, lead to overall cross-linguistic differences in reliance on body-rime correspondences.

In English, two different corpus analyses have shown considerable gains in taking account body information in contrast to single vowel graphemes (in monosyllabic words with a CVC structure). One analysis showed an increase in predictability from 62% to 80% (Treiman et al., 1995), and another and increase from 48% to 91% (Peereman & Content, 1998). In German, Ziegler et al. (2000) report that even when only single-letter and context-insensitive grapheme-phoneme correspondences are implemented in the DRC, 88.5% of all monosyllabic words can be read correctly by the sublexical route. In the more specific case of vowel pronunciations, a context-insensitive rule, that all single vowel letters are short, would result in 77.9% correct vowel pronunciations for all monosyllabic words (Perry, Ziegler, Braun, et al., 2010).

Compared to the English consistency estimates of around 48-62%, it is clear that even relying on simple rules, there is already little uncertainty in a word's pronunciation in German. This suggests that the gains of taking into account body-rime correspondences, in contrast to English, should be relatively smaller (Ziegler &

Goswami, 2005). Thus, finding reliance on bodies in German *per se* does not provide evidence against the notion that bodies are psychologically salient because they reduce unpredictability. However, the fact that we do not find any cross-linguistic differences is inconsistent with this view. In the following sections, we consider alternative accounts that could explain why reliance on bodies develops.

Phonological saliency of rimes and its role in reliance on bodies during reading

Originally, the large-versus-small units first debate was framed in terms of the importance of phonological awareness of phonemes versus rimes as predictors of reading ability (Duncan et al., 1997; Goswami, 1991, 1999, 2002; Goswami & Bryant, 1990; Hulme et al., 2002; Treiman et al., 1990). The strong correlation between phonological awareness and reading ability is one of the best-established findings in the literature on reading acquisition (e.g., see Castles & Coltheart, 2004; Snowling, 2000; Ziegler & Goswami, 2005, for reviews).

It is intuitive that sublexical print-to-speech correspondences, or establishing the link between an orthographic and a phonological unit, depends on the ability to perceive the phonological unit. Here, we get to what Ziegler and Goswami (2005) refer to as the "availability" problem. Mostly, children beginning to learn to read in alphabetic orthographies are taught to map letters onto their corresponding sounds (i.e., phonemes). Yet, it is well established that reliance on "larger" phonological units, such as syllables, emerges before awareness of phonemes (e.g., Goswami & Bryant, 1990; Liberman, Shankweiler, Fischer, & Carter, 1974; Treiman & Zukowski, 1991; Ziegler & Goswami, 2005). Furthermore, evidence suggests that explicit awareness of phonemes emerges as a result of reading instruction, and more specifically instructional methods that emphasise the role of phonemes (Alegria, Pignot, & Morais, 1982; Mann, 1986; Morais, Alegria, & Content, 1987; Read, Zhang, Nie, & Ding, 1986; Wimmer, Landerl, Linortner, &

Hummer, 1991). Therefore, young children are on the one hand taught to read by using "small" units, while on the other hand showing little evidence for explicit awareness of these.

Based on these observations, one might expect that young children start learning to read by relying on large orthographic clusters, and develop increasingly refined sensitivity to smaller orthographic units (as their phonemic awareness develops) as a result of reading exposure (Goswami, 1993). In Papers 3 and 5, we set out to test this prediction. Although the results from Paper 3 were inconclusive - as we found stability of reliance on bodies across the age groups that were tested - a different pattern emerged in Paper 5 (for the German children - unfortunately, the English data were too messy to draw any conclusions about the developmental trajectory across age). Here, we found that for German children, the reliance on context-insensitive grapheme-phoneme correspondences decreases with age, while the reliance on body-rime correspondences increases.

How do we explain this discrepancy between the theory and the results? At this stage, it is important to closely consider the link between phonological awareness and reading. As has been argued previously (Castles & Coltheart, 2004; Jackson & Coltheart, 2001), phoneme awareness is a distal skill, and not part of the reading system itself: although it supports various skills that are important for reading, the exact relationship between phonological awareness and reading is still unclear. Goswami and Bryant argue that "if children are aware of onsets and rimes and connect these intra-syllabic speech units to writing, they must be making connections between sounds and whole letter sequences of letters." (p. 19, Goswami & Bryant, 1990). A questionable link in this chain is the relationship between awareness of onsets and rimes, and connecting these to writing. Although phonological awareness of onsets and rimes may facilitate reading

acquisition (Bryant, Bradley, Maclean, & Crossland, 1989; Bryant, Maclean, Bradley, & Crossland, 1990), our data indicate that the link between rimes and their orthographic equivalents becomes more important as the children become older, suggesting that these links are only established as a result of reading exposure.

Therefore, although initial phonological awareness is beyond doubt important in learning to read, the phonological awareness of a specific type of unit does not seem to directly translate to the establishing of print-speech correspondences. Future research could focus on establishing the exact chain of events that (indirectly) link phonological awareness to the psychological reality of orthography-phonology correspondences for specific types of units.

The development of speed

The final pressure that might encourage the sublexical route to rely on body units is the development of reading speed. A system that decodes, say, the word *stop* in a letter-by-letter fashion will be two times slower than a system that decodes the same word by its onset-rime structure, provided that letter clusters can be learned to be recognised as the same speed as single letters.

A previous theory of reading has proposed that reliance on large units, such as syllables, develops in order to increase reading speed (LaBerge & Samuels, 1974), but this claim has only recently been put under thorough methodological scrutiny. Several training studies have devised training programs for poor readers which emphasise the role of larger-than-grapheme units in order to increase reading fluency. Aside from the practical utility of these experiments, they provide a strong test of the proposal that relying on larger clusters causes faster reading speed. Such studies have been conducted mostly in shallow orthographies, as here deficient reading speed is the prominent feature of dyslexia (Wimmer, 1993). Overall, these studies show benefits from training children

by emphasising the importance of syllables (Ecalle, Magnan, & Calmus, 2009; Huemer, Aro, Landerl, & Lyytinen, 2010; Tressoldi, Vio, & Iozzino, 2007; Wentink, VanBon, & Schreuder, 1997), but little benefit, beyond explicitly trained words, of training subsyllabic units (Das Smaal, Klapwijk, & van der Leij, 1996; Marinus, de Jong, & van der Leij, 2012; Thaler, Ebner, Wimmer, & Landerl, 2004).

It is unclear how to interpret these findings from our perspective: it appears that syllables and other subsyllabic clusters may have different roles in speeding up word recognition. Therefore, it is not clear that relying on larger units increases speed, but it has also been pointed out that it is difficult to train children to rely on larger units (Marinus et al., 2012).

Summary

In this section, we considered the *how* and *why* of body-rime correspondences. Although we showed that both German and English readers rely on bodies (Papers 1 - 5), we can provide only speculative answers to how the cognitive system chooses to rely on a specific orthographic unit in a given circumstance, and why reliance on bodies develops in the first place.

Methodological challenges for cross-linguistic research

Matching items across languages

In conducting any experiment, it is important to rule out alternative explanations that, in addition to the manipulation, could account for the results. In psycholinguistic research, this is especially complicated due to the inter-correlated nature of most linguistic constructs (Andrews, 1997; Cutler, 1981; Kliegl et al., 2004; Yap, Balota, Sibley, & Ratcliff, 2012): choosing random items that are either high or low one psycholinguistic variable usually means that they differ on a wide range of other item

characteristics. Psycholinguistic experiments often use orthogonal designs, where the items that are chosen need to be de-correlated with other linguistic constructs. In cross-linguistic reading research, these issues become even more pronounced, because of the systematic differences in language characteristics (see Papers 2 and 6).

In English and German, several studies have used cognates in an attempt to circumvent this problem (Frith et al., 1998; Landerl et al., 1997; Rau et al., 2015; Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003). This method draws on the presence of words in English and German which have the same spelling and meaning in both languages, such as the word *Park*. The rationale behind the cognate design is that finding a behavioural difference suggests that English and German readers process these "identical" items in a different way.

However, there are problems with the cognate design. Firstly, it not the case that cognates are identical across languages. The word *Bank*, for example, is a cognate, but it has an additional meaning in German (*bench*). Cognates like *baseball* and *bratwurst* are obvious examples of words that occur more frequently in one language than the other. The word *yacht*[28] is a perfect illustration that even if the lexical-semantic variables are approximately matched, there may be substantial differences in lexical-orthographic properties: In English, this word is notoriously irregular, and a "hermit", as it has no body neighbours. In German, its pronunciation is perfectly predictable by the use of relatively simple letter-sound correspondence rules, and it comes from a dense orthographic neighbourhood as it has 15 body neighbours (*Macht*, *Fracht*, etc.).

Secondly, the use of cognates does not solve the problem of controlling for orthographic characteristics other than the concept of interest, especially if these vary systematically across languages. Thus, cognate designs are susceptible to the same

[28] Both "Jacht" and "Yacht" are legitimate spellings of this word in German.

problems as any other cross-linguistic comparison. Importantly, lexical and sublexical characteristics that co-vary with linguistic properties of the orthography may lead to systematic differences in the item sets across languages, even if only cognates are used. English and German differ from each other in terms of orthographic depth, but also in terms of orthographic density (Marinus, Nation, & de Jong, under review). As discussed in Papers 2 - 3, this confound becomes critical when we interpret the experiments which have provided key evidence for the assumption underlying the Psycholinguistic Grain Size Theory (Ziegler et al., 2001; Ziegler, Perry, Ma-Wyatt, et al., 2003).

Matching participants across languages

Once a well controlled item-set has been created, the next set of issues relates to the matching of participants across languages. Here, we discuss three systematic confounds that are associated with matching participants across orthographies varying in orthographic depth, and specifically between English and German: (1) reading instruction methods, (2) cultural differences in reading instruction onset, and (3) differential reading profiles as a function of orthographic depth.

Reading instruction methods

Reading instruction methods, and specifically the preference towards whole-word or phonics instruction, differ systematically as a function of orthographic depth. In Australian primary schools, for example, a whole-word approach to reading has been commonly used throughout the last decades (de Lemos, 2002). Teaching children to read in a deep orthography seems to evoke the illusion that teaching phonics, or print-to-speech correspondences, is unhelpful, because these correspondences are unreliable. In Germany, in contrast, it is customary to teach children to read via phonics instruction (Landerl, 2000). Therefore, instruction method is a systematic confound that is associated with orthographic depth.

This is important to take into account, particularly for cross-linguistic studies on the nature of sublexical processing: A recent study has shown that even in skilled adult readers of English, the sublexical mechanisms used during nonword reading are different depending on whether the participants had learned to read via phonics or whole-word instruction methods in childhood (Thompson et al., 2009). Therefore, unless a cross-linguistic study has controlled for this confound, it is impossible to establish whether behavioural differences in the nature of sublexical processing are due to reading instruction methods, or due to characteristics of the orthography.

For adults, little attention has been paid to this issue, as it is assumed that undergraduate students of German and English-speaking universities are comparable in terms of their overall reading profiles. In line with this assumption, we have not considered systematic differences across countries in the adult population during recruitment or testing. This may be a drawback of the studies presented in this thesis.

In Papers 3 and 4, we attempted to circumvent this issue by using bilingual participants. Finding language interactions while using the same participants reading in different orthographies shows that the cross-linguistic differences are attributable to characteristics of the orthography, rather than individual differences (Frost, Kugler, Deutsch, & Forster, 2005). Finding no differences in the same participants reading in a different language, conversely, suggests that positive findings of cross-linguistic differences in previous studies using a between-subject design may have been due to random or systematic individual variability. In Paper 3, the participants were children from a bilingual school in Australia. Here, we found no interaction between language and body-N (a marker effect of reliance on large units), suggesting that sublexical processes underlying reading in English and German may be less different than previously assumed.

In Paper 4, we tested adult German and English native speakers in their respective language, but also compared German native speakers who lived in Australia both to German native speakers who lived in Germany (for the German items) and to Australian monolingual participants (reading English items). We found little difference between the German/English bilinguals and the monolingual participants – as the German participants would almost certainly have received phonics instruction, this raises our confidence in concluding that the results of the study, and the cross-linguistic differences in the overall patterns, are not entirely due to reading instruction. In Paper 5, we applied the same procedure to German and English monolingual children. As we did not collect data with bilingual children on this task, reading instruction is a potential confound. However, as for the English items in Paper 4, we found little difference between monolingual English and bilingual German native speakers, so it is unlikely that the cross-linguistic differences can be fully explained by differences in reading instruction.

Onset of reading instruction

An additional instruction-related confound is the onset of reading instruction. Australian children start pre-school, which includes some formal reading instruction, at age 5. In contrast, German children start school at age 6 or 7, and it is uncommon for German kindergartens or parents to teach children to read before they start school. The difference between the age of onset of reading instruction is a systematic confound, which is related to orthographic depth: as it takes longer to read in a deep language, reading instruction tends to start at an earlier age than in shallow orthographies.

Choosing appropriate tests: Differential reading profiles in English and German

The issue of choosing an appropriate test for matching children on their reading ability has been debated in detail in relation to a reading-age match design, which is commonly used in studies on dyslexia (e.g., Jackson & Coltheart, 2001). Here, children

are matched based on their reading ability, rather than their age which, in the case of dyslexia research, is designed to ensure that any differences between the two groups are due to a deficit that is specific to dyslexia, and not due to reading proficiency. Matching children on their reading ability, however, requires an important decision: namely, what particular reading test they should be matched on.

In the case of cross-linguistic research, it has been shown that English children lag behind German children on nonword reading to a greater extent than on their real word reading ability (Frith et al., 1998; Landerl, 2000; Seymour et al., 2003). Therefore, matching children on nonword reading ability across languages will mean that the English children are relatively better at word reading, and matching children on real words will mean that the German children are relatively better at nonword reading. Furthermore, there are some studies indicating that given their word reading ability, English children are better at text comprehension compared to children learning to read in Welsh, a shallow orthography (Ellis & Hooper, 2001; Hanley, Masterson, Spencer, & Evans, 2004). If this reflects a systematic difference that is associated with orthographic depth (as might be suggested by the Orthographic Depth Hypothesis, where lexical-semantic processing should be relatively more important in English compared to shallow orthographies), then matching children on their reading aloud accuracy would also result in systematic differences in terms of reading comprehension.

There is no straightforward solution to this problem. In Paper 3, we use bilingual children for a cross-linguistic comparison. However, the same child can be better at one language compared to the other: in fact, even though the children in our study attended the same school, exposure to one language over the other at home differed across individuals. Therefore, this design does not guarantee that the children are, in any way, matched for their reading ability.

In Paper 5, the participants are monolingual children speaking either German or English. Here, the Australian sample happened to consist of relatively good readers for their age. As a result, the Australian children showed better speeded sight word reading ability than the German children. In more difficult tests designed to assess the efficiency of the lexical or sublexical route, however, the German children outperformed the Australian children. Given the across-language discrepancy between sight word reading ability and the performance on those more complex tasks, we conclude that both the development of lexical and sublexical processing appears to lag in English due to the overall complexity of sublexical correspondences.

Summary

Orthographic depth co-varies with several cultural practices that create potential participant-level confounds that need to be considered in cross-linguistic research. Whole-word instruction methods, as opposed to phonics (letter-sound) teaching, are particularly popular in countries with deep orthographies, because it may seem intuitive that there is no point in teaching letter-sound correspondences in an orthography where these appear, at first glance, to be more misleading than helpful. Instruction method at school, in turn, has been shown to affect participants' sublexical processes, even in adulthood (Thompson et al., 2009).

An additional systematic confound is the onset of reading instruction: schooling starts earlier in English- compared to German-speaking countries, as it takes longer to read in English compared to German (Seymour et al., 2003). Furthermore, there is evidence that the developmental trajectory of the reading profile, or the relative proficiency of various subskills underlying reading, differs as a function of orthographic depth (e.g., Ellis & Hooper, 2001; Seymour et al., 2003; Wimmer & Goswami, 1994). These issues need to be considered in matching participants across languages. With

children, a stringent match may be impossible due to differential patterns of development across languages. As a result, one needs to consider to what extent differences between the two samples may compromise the conclusions about behavioural differences related to characteristics of the orthography that can be drawn from the experiment.

Conclusions

In summary, we have shown that conducting cross-linguistic research requires a careful consideration of methodological and theoretical issues, and that the results of a study should be interpreted in the context its of methodological and theoretical limitations. This is true for all psycholinguistic research, but perhaps it is particularly important for cross-linguistic comparisons, because languages differ from each other on many attributes, therefore it is not always clear what language-level difference drives a behavioural difference.

In thesis, we have addressed some issues relating to cross-linguistic reading research. The major findings can be summarised as follows: (1) Body-rime correspondences are used by German and English readers to the same extent, (2) context-sensitive correspondences are more important for English than German, (3) the development of the sublexical route follows a small-to-large unit trajectory, where older children acquire body-rime and context-sensitive correspondences as a result of reading exposure.

We have also isolated a range of questions that need to be addressed by future research. The finding that the sublexical route processes different sublexical units in parallel provides a benchmark for computational models of reading, but does not provide a specific computational mechanism that would explain how conflicts between different units are resolved. The finding that both German and English readers rely on body-rime correspondences to the same extent raises the question as to why the reliance on these

correspondences emerges, as existing theories on the reliance on large sublexical units predict cross-linguistic differences (Goswami, 1999; Ziegler & Goswami, 2005). Finally, future research is needed to determine how different aspects of orthographic depth contribute to cross-linguistic differences in the cognitive mechanisms underlying reading.

Appendix: Ethics approval (Macquarie University and Potsdam University)

MACQUARIE UNIVERSITY

Xenia Schmalz <xenia.schmalz@mq.edu.au>

RE: Ethics Application - Final Approval (Subject to Condition/s) (Ref. No.5201200053D)

7 messages

Fhs Ethics <fhs.ethics@mq.edu.au> Tue, Apr 10, 2012 at 9:59 AM
To: Prof Anne Castles <anne.castles@mq.edu.au>
Cc: Dr Eva Marinus <eva.marinus@mq.edu.au>, Ms Xenia Schmalz <xenia.schmalz@students.mq.edu.au>

Dear Prof Castles,

RE: 'Flexibility of cognitive processing in skilled word recognition and learning to read ' (Ref: 5201200053)

Thank you for your recent correspondence. Your response has addressed the issues raised by the Faculty of Human Sciences Human Research Ethics Sub-Committee. Approval of this application has been granted and you may now proceed with your research.

This approval is subject to the following conditions:

1. Amendment request should be submitted if the researchers decide to proceed with data collection from German participants;
2. A copy of emails or letters from the principals should be provided when they are available.

This research meets the requirements of the National Statement on Ethical Conduct in Human Research (2007). The National Statement is available at the following web site:

http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf.

The following personnel are authorised to conduct this research:

Chief Investigator: Professor Anne Castles

Other Personnel:
Dr Eva Marinus
Ms Xenia Schmalz

NB. STUDENTS: IT IS YOUR RESPONSIBILITY TO KEEP A COPY OF THIS APPROVAL EMAIL TO SUBMIT WITH YOUR THESIS.

Universität Potsdam · Am Neuen Palais 10 · 14469 Potsdam

Ethikkommission
Vorsitzender
Prof. Dr. Esser

Telefon: (03 31) 9 77 17 91
Telefax: (03 31) 9 77 10 89
Datum: 30.5.2013

Frau Xenia Schmalz
Allg. Psych. I (Abt. Prof. Kliegl)
Department Psychologie

Endbescheid - Antrag Nr. 4/2013

Sehr geehrte Frau Schmalz,

die Ethikkommission erhebt keine Einwände gegen das Forschungsprojekt

„Cross-linguistic differences in the development of lexical and sublexical strategies in reading (Sprachübergreifende Unterschiede in der Entwicklung lexikalischer und sublexikalischer Strategien im Lesen)".

Die Anmerkungen der Ethikkommission wurden hinreichend bearbeitet bzw. umgesetzt.

Ich wünsche Ihnen für die Durchführung Ihres Projektes viel Erfolg.

Freundliche Grüße

Prof. Dr. Esser
Vorsitzender der Ethikkommission

Prof. Dr. Günter Esser
Professur Klinische Psychologie / Psychotherapie
UNIVERSITÄT POTSDAM
Karl-Liebknecht-Str. 24 - 25
14476 Potsdam, OT Golm

References

Alegria, J., Pignot, E., & Morais, J. (1982). Phonetic Analysis of Speech and Memory Codes in Beginning Readers. *Memory & Cognition, 10*(5), 451-456. doi: 10.3758/Bf03197647

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

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*(4), 439-461.

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

Andrews, S., Woollams, A., & Bond, R. (2005). Spelling-sound typicality only affects words with digraphs: Further qualifications to the generality of the regularity effect on word naming. *Journal of Memory and Language, 53*(4), 567-593. doi: 10.1016/J.Jml.2005.04.002

Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language, 63*(2), 180-196. doi: 10.1016/J.Jml.2010.03.005

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

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database. Release 2 (CD-ROM): Linguistic Data Consortium, University of Pennsylvania.

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: Doi 10.3758/Bf03193014

Besner, D., & Smith, M. (1992). Basic Processes in Reading: Is the Orthographic Depth Hypothesis Sinking? In R. Frost & L. Katz (Eds.), *Orthography, Phonology, Morphology, and Meaning* (pp. 45-66). Amsterdam: Elsevier Science Publishers.

Borgwaldt, S., Hellwig, F., & de Groot, A. (2004). Word-initial entropy in five languages: Letter to sound, and sound to letter. *Written Language & Literacy, 7*(2), 165-184.

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

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

Brown, G., & Deavers, R. (1999). Units of Analysis in Nonword Reading: Evidence from Children and Adults. *Journal of Experimental Child Psychology, 73*, 208-242.

Bruck, M., Genesee, F., & Caravolas, M. (1997). A cross-linguistic study of early literacy acquisition. *Foundations of Reading Acquisition and Dyslexia*, 145-162.

Bryant, P. E., Bradley, L., Maclean, M., & Crossland, J. (1989). Nursery Rhymes, Phonological Skills and Reading. *Journal of Child Language, 16*(2), 407-428.

Bryant, P. E., Maclean, M., Bradley, L. L., & Crossland, J. (1990). Rhyme and Alliteration, Phoneme Detection, and Learning to Read. *Developmental Psychology, 26*(3), 429-438. doi: 10.1037/0012-1649.26.3.429

Burani, C., & Arduino, L. S. (2004). Stress regularity or consistency? Reading aloud Italian polysyllables with different stress patterns. *Brain and Language, 90*, 318-325. doi: 10.1016/s0093-934x(03)00444-9

Byrd, R., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing, 16*, 1190-1208.

Campbell, R., & Besner, D. (1981). This and thap - constraints on the pronunciation of new, written words. *Quarterly Journal of Experimental Psychology, 33A*, 375-396.

Caravolas, M., & Bruck, M. (1993). The Effect of Oral and Written Language Input on Childrens Phonological Awareness - a Cross-Linguistic Study. *Journal of Experimental Child Psychology, 55*(1), 1-30. doi: Doi 10.1006/Jecp.1993.1001

Caravolas, M., Leråg, A., Mousikou, P., Efrim, C., Litavsky, M., Onochie-Quintanilla, E., . . . Hulme, C. (2012). Common Patterns of Prediction of Literacy Development in Different Alphabetic Orthographies. *Psychological Science, 23*(6), 678-686. doi: 10.1177/0956797611434536

Caravolas, M., Lervag, A., Defior, S., Malkova, G. S., & Hulme, C. (2013). Different Patterns, but Equivalent Predictors, of Growth in Reading in Consistent and Inconsistent Orthographies. *Psychological Science, 24*(8), 1398-1407. doi: 10.1177/0956797612473122

Caravolas, M., Volin, J., & Hulme, C. (2005). Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from Czech and English children. *Journal of Experimental Child Psychology, 92*(2), 107-139. doi: 10.1016/j.jecp.2005.04.003

Cassidy, S., & Harrington, J. (2001). Multi-level annotation in the Emu speech database management system. *Speech Communication, 33*(1-2), 61-77. doi: 10.1016/S0167-6393(00)00069-8

Castles, A., & Coltheart, M. (1993). Varieties of developmental dyslexia. *Cognition, 47*, 149-180.

Castles, A., & Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition, 91*, 77-111.

Castles, A., Coltheart, M., Larsen, L., Jones, P., Saunders, S., & McArthur, G. (2009). Assessing the basic components of reading: A revision of the Castles and Coltheart test with new norms. *Australian Journal of Learning Difficulties, 14*(1), 67-88. doi: 10.1080/19404150902783435

Chang, W., Maries, A., & Perfetti, C. (2014, 17/7/2014). *Visual Orthographic Variation across Writing Systems.* Paper presented at the Scientific Studies of Reading SSSR Santa Fe.

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

Colombo, L. (1992). Lexical stress effect and its interaction with frequency in word pronunciation. *Journal of Experimental Psychology-Human Perception and Performance, 18*(4), 987-1003. doi: 10.1037//0096-1523.18.4.987

Coltheart, M. (2012). Dual-route theories of reading aloud. In J. Adelman (Ed.), *Visual Word Recognition*. Hove, UK: Psychology Press.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of Reading Aloud - Dual-Route and Parallel-Distributed-Processing Approaches. *Psychological Review, 100*(4), 589-608.

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance, VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

Coltheart, M., & Rastle, K. (1994). Serial Processing in Reading Aloud: Evidence for Dual-Route Models of Reading. *Journal of Experimental Psychology: Human Perception & Performance, 20*(6), 1197-1211.

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

Coltheart, V., & Leahy, J. (1992). Children's and Adults' Reading of Nonwords: Effects of Regularity and Consistency. *Journal of Experimental Psychology: Learning, Memory & Cognition, 18*(4), 718-729.

Content, A. (1991). The effect of spelling-to-sound regularity on naming in French. *Psychological Research, 53*(3), 3-12.

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

Cox, F., & Palethorpe, S. (2007). Australian English. *Journal of the International Phonetic Association, 37*, 341-350.

Cui, L., Drieghe, D., Yan, G., Bai, X., Chi, H., & Liversedge, S. (2012). Parafoveal processing across different lexical constituents in Chinese reading. *The Quarterly Journal of Experimental Psychology*, 1-14. doi: 10.1080/17470218.2012.720265

Cummine, J., Amyotte, J., Pancheshen, B., & Chouinard, B. (2011). Evidence for the Modulation of Sub-Lexical Processing in Go No-Go Naming: The Elimination of the Frequency x Regularity Interaction. *Journal of Psycholinguistic Research, 40*(5-6), 367-378. doi: 10.1007/s10936-011-9174-2

Cutler, A. (1981). Making up materials is a confounded nuisance: or Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition, 10*(1-3), 65-70. doi: 10.1016/0010-0277(81)90026-3

Das Smaal, E. A., Klapwijk, M. J. G., & van der Leij, A. (1996). Training of perceptual unit processing in children with a reading disability. *Cognition and Instruction, 14*(2), 221-250. doi: 10.1207/S1532690xci1402_3

de Jong, P. (2006). Units and routes of reading in Dutch. *Developmental Science, 9*(5), 441-442.

de Lemos, M. (2002). Closing the gap between research and practice: Foundations for the acquisition of literacy. *Literacy and Numeracy Research Literature Reviews.* from http://research.acer.edu.au/literacy_numeracy_reviews/1

Dehaene, S. (2009). *Reading in the brain: The new science of how we read*. London: Penguin.

Duncan, L., Castro, S. L., Defior, S., Seymour, P. H. K., Baillie, S., Leybaert, J., . . . Serrano, F. (2013). Phonological development in relation to native language and literacy: Variations on a theme in six alphabetic orthographies. *Cognition, 127*(3), 398-419. doi: 10.1016/J.Cognition.2013.02.009

Duncan, L., Seymour, P., & Hill, S. (1997). How important are rhyme and analogy in beginning reading? *Cognition, 63*, 171-208.

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods Instruments & Computers, 36*(3), 488-499. doi: 10.3758/bf03195595

Ecalle, J., Magnan, A., & Calmus, C. (2009). Lasting effects on literacy skills with a computer-assisted learning using syllabic units in low-progress readers. *Computers & Education, 52*(3), 554-561. doi: 10.1016/J.Compedu.2008.10.010

Ellis, N., & Hooper, M. (2001). Why learning to read is easier in Welsh than in English: Orthographic transparency effects evinced with frequency-matched tests. *Applied Psycholinguistics, 22*, 571-599.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125*(6), 777-799. doi: Doi 10.1037//0033-2909.125.6.777

Feldman, L. B., & Turvey, M. T. (1983). Word Recognition in Serbo-Croatian Is Phonologically Analytic. *Journal of Experimental Psychology-Human Perception and Performance, 9*(2), 288-298. doi: 10.1037//0096-1523.9.2.288

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

Forster, K. I., & Taft, M. (1994). Bodies, Antibodies, and Neighborhood-Density Effects in Masked Form Priming. *Journal of Experimental Psychology-Learning Memory and Cognition, 20*(4), 844-863. doi: Doi 10.1037//0278-7393.20.4.844

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

Frost, R. (2012b). A universal approach to modeling visual word recognition and reading: Not only possible, but also inevitable Response. *Behavioral and Brain Sciences, 35*(5), 310-329. doi: 10.1017/S0140525x12000635

Frost, R., & Bentin, S. (1992). Reading consonants and guessing vowels: Visual word recognition in Hebrew orthography. *Advances in psychology, 94*, 27-44.

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.

Frost, R., Kugler, T., Deutsch, A., & Forster, K. I. (2005). Orthographic structure versus morphological structure: Principles of lexical organization in a given language. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(6), 1293-1326. doi: 10.1037/0278-7393.31.6.1293

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.

Goswami, U. (1991). Learning about Spelling Sequences: The Role of Onsets and Rimes in Analogies in Reading. *Child Development, 62*, 1110-1123.

Goswami, U. (1993). Toward an Interactive Analogy Model of Reading Development - Decoding Vowel Graphemes in Beginning Reading. *Journal of Experimental Child Psychology, 56*(3), 443-475. doi: Doi 10.1006/Jecp.1993.1044

Goswami, U. (1998). The Role of Analogies in the Development of Word Recognition. In J. Metsala & L. Ehri (Eds.), *Word Recognition in Beginning Literacy*.

Goswami, U. (1999). The relationship between phonological awareness and orthographic representations in different orthographies. In M. Harris & G. Hatano (Eds.), *Learning to read and write: A cross-linguistic perspective* (pp. 134-156). Cambridge: Cambridge Press.

Goswami, U. (2002). In the Beginning was the Rhyme? A Reflection on Hulme, Hatcher, Nation, Brown, Adams and Stuart (2002). *Journal of Experimental Child Psychology, 82*, 47-57.

Goswami, U., & Bryant, P. (1990). *Phonological Skills and Learning to Read*. Hove: Lawrence Erlbaum Associates Ltd.

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

Goswami, U., Porpodas, C., & Wheelwright, S. (1997). Children's orthographic representations in English and Greek. *European Journal of Psychology of Education, 12*(3), 273-292.

Goswami, U., & Ziegler, J. (2006). Fluency, phonology and morphology: a response to the commentaries on becoming literate in different languages. *Developmental Science, 9*(5), 451-453.

Goswami, U., Ziegler, J., Dalton, L., & Schneider, W. (2003). Nonword reading across orthographies: How flexible is the choice of reading units? *Applied Psycholinguistics, 24*, 235-247. doi: 10.1017.S0142716403000134

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *Journal of Statistical Software, 33*, 1-33.

Hanley, J. R., Masterson, J., Spencer, L. H., & Evans, D. (2004). How long do the advantages of learning to read a transparent orthography last? An investigation of the reading skills and reading impairment of Welsh children at 10 years of age. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 57*(8), 1393-1410. doi: 10.1080/02724980343000819

Harm, M., & Seidenberg, M. (1999). Phonology, Reading Acquisition, and Dyslexia: Insights From Connectionist Models. *Psychological Review, 106*(3), 491-528.

Harm, M., & Seidenberg, M. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*(3), 662-720. doi: 10.1037/0033-295x.111.3.662

Havelka, J., & Rastle, K. (2005). The assembly of phonology from print is serial and subject to strategic control: Evidence from Serbian. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(1), 148-158. doi: 10.1037/0278-7393.31.1.148

Hino, Y., & Lupker, S. J. (2000). Effects of word frequency and spelling-to-sound regularity in naming with and without preceding lexical decision. *Journal of Experimental Psychology-Human Perception and Performance, 26*(1), 166-183. doi: 10.1037//0096-1523.26.1.166

Huang, H. S., & Hanley, J. R. (1995). Phonological Awareness and Visual Skills in Learning to Read Chinese and English. *Cognition, 54*(1), 73-98. doi: 10.1016/0010-0277(94)00641-W

Huemer, S., Aro, M., Landerl, K., & Lyytinen, H. (2010). Repeated Reading of Syllables Among Finnish-Speaking Children With Poor Reading Skills. *Scientific Studies of Reading, 14*(4), 317-340. doi: 10.1080/10888430903150659

Huey, E. B. (1908). *The psychology and pedagogy of reading*: The Macmillan Company.

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The Causal Role of Phoneme Awareness and Letter-Sound Knowledge in Learning to

Read: Combining Intervention Studies With Mediation Analyses. *Psychological Science, 23*(6), 572-577. doi: 10.1177/0956797611435921
Hulme, C., Hatcher, P. J., Nation, K., Brown, A., Adams, J., & Stuart, G. (2002). Phoneme awareness is a better predictor of early reading skill than onset-rime awareness. *Journal of Experimental Child Psychology, 82*(1), 2-28. doi: 10.1006/jecp.2002.2670
Hutzler, F., Ziegler, J., Perry, C., Wimmer, H., & Zorzi, M. (2004). Do current connectionist learning models account for reading development in different languages? *Cognition, 91*, 273-296.
Jackson, N., & Coltheart, M. (2001). *Routes to reading success and failure: Toward an integrated cognitive psychology of atypical reading.* New York, NY: Psychology Press.
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. S. (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
Job, R., Satori, G., Masterson, J., & Coltheart, M. (1984). Developmental Surface Dyslexia in Italian. In R. Malatesha & H. Whitaker (Eds.), *Dyslexia: A global issue* (pp. 133-141). Hague: Martinus Nijhoff.
Jouravlev, O., & Lupker, S. J. (2014). Stress consistency and stress regularity effects in Russian. *Language Cognition and Neuroscience, 29*(5), 605-619. doi: 10.1080/01690965.2013.813562
Justi, C., & Justi, F. (2009). The Effects of Lexicality, Frequency, and Regularity in Brazilian Portuguese Speaking Children [Portuguese]. *Psicologia: Reflexão e Critica, 22*(2), 163-172.
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.
Kay, J., & Marcel, A. (1981). One process, not two, in reading aloud: Lexical analogies do the work of non-lexical rules. *The Quarterly Journal of Experimental Psychology Section A, 33*(4), 397-413. doi: 10.1080/14640748108400800
Kerek, E., & Niemi, P. (2012). Grain-size units of phonological awareness among Russian first graders. *Written Language & Literacy, 15*(1), 80-113. doi: http://dx.doi.org/10.1075/wll.15.1.05ker
Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627-633. doi: 10.3758/brm.42.3.627
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: Doi 10.3758/S13428-011-0118-4
Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*(1-2), 262-284. doi: Doi 10.1080/09541440340000213
Kuo, W. J., Yeh, T. C., Lee, J. R., Chen, L. F., Lee, P. L., Chen, S. S., . . . Hsieh, J. C. (2004). Orthographic and phonological processing of Chinese characters: an fMRI

study. *NeuroImage, 21*(4), 1721-1731. doi: Doi 10.1016/J.Neuroimage.2003.12.007
LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293-323.
Landerl, K. (2000). Influences of orthographic consistency and reading instruction on the development of nonword reading skills. *European Journal of Psychology of Education, 15*(3), 239-257.
Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppanen, P. H. T., Lohvansuu, K., . . . Schulte-Korne, G. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry, 54*(6), 686-694. doi: 10.1111/Jcpp.12029
Landerl, K., & Reitsma, P. (2005). Phonological and morphological consistency in the acquisition of vowel duration spelling in Dutch and German. *Journal of Experimental Child Psychology, 92*(4), 322-344. doi: 10.1016/J.Jecp.2005.04.005
Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition, 63*, 315-334.
Liberman, I., Liberman, A., Mattingly, I., & Shankweiler, D. (1980). Orthography and the beginning reader. In J. Kavenagh & R. L. Venezky (Eds.), *Orthography, Reading, and Dyslexia*. Baltimore: University Park Press.
Liberman, I., Shankweiler, D., Fischer, F., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology, 18*(2), 201-212.
Mann, V. (1986). Phonological awareness: The role of reading experience. *Cognition, 24*, 65-92.
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., & van der Leij, A. (2012). Increasing Word-Reading Speed in Poor Readers: No Additional Benefits of Explicit Letter-Cluster Training. *Scientific Studies of Reading, 16*(2), 166-185. doi: 10.1080/10888438.2011.554471
Marinus, E., & de Jong, P. F. (2011). Dyslexic and typical-reading children use vowel digraphs as perceptual units in reading. *Quarterly Journal of Experimental Psychology, 64*(3), 504-516. doi: 10.1080/17470218.2010.509803
Marinus, E., Kohnen, S., & McArthur, G. (2013). Australian comparison data for the Test of Word Reading Efficiency (TOWRE). *Australian Journal of Learning Difficulties, 18*(2), 199-212.
Masterson, J., Stuart, M., Dixon, M., Lovejoy, D., & Lovejoy, S. (2003). The Children's Printed Word Database., from University of Essex http://www.essex.ac.uk/psychology/cpwd/
McBride-Chang, C., Chen, H. C., Kasisopa, B., Burnham, D., Reilly, R., & Leppanen, P. (2012). What and where is the word? *Behavioral and Brain Sciences, 35*(5), 295-296. doi: 10.1017/S0140525x1200009x
Metsala, J. L., Stanovich, K. E., & Brown, G. D. A. (1998). Regularity effects and the phonological deficit model of reading disabilities: A meta-analytic review. *Journal of Educational Psychology, 90*(2), 279-293. doi: 10.1037/0022-0663.90.2.279
Moll, K., & Landerl, K. (2010). SLRT-II: Lese- und Rechtschreibtest; Weiterentwicklyng des Salzburger Lese- und Rechtschreibtests (SLRT): Huber.

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., . . . Landerl, K. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learning and Instruction, 29*, 65-77. doi: 10.1016/J.Learninstruc.2013.09.003

Monfort, A. (1995). *Statistics and economic models* (Vol. 2). Cambridge: Cambridge University Press.

Morais, J., Alegria, J., & Content, A. (1987). The relationship between segmental analysis and alphabetic literacy: An interactive view. *Cahiers de Psychologie Cognitive. European Bulletin of Cognitive Psychology, 7*, 415-438.

Morey, R. D., & Rouder, J. N. (2014). Package "BayesFactor". Retrieved 9.8.2014, from http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

Nation, K., Allen, R., & Hulme, C. (2001). The limitations of orthographic analogy in early reading development: Performance on the clue-word task depends on phonological priming and elementary decoding skill, not the use of orthographic analogy. *Journal of Experimental Child Psychology, 80*(1), 75-94. doi: 10.1006/jecp.2000.2614

Paap, K. R., & Noel, R. W. (1991). Dual-Route Models of Print to Sound - Still a Good Horse Race. *Psychological Research-Psychologische Forschung, 53*(1), 13-24. doi: Doi 10.1007/Bf00867328

Parkin, A., McMullen, M., & Graystone, D. (1986). Spelling-to-sound regularity affects pronunciation latency but not lexical decision. *Psychological Research, 48*, 87-92.

Patterson, K., & Behrmann, M. (1997). Frequency and consistency effects in a pure surface dyslexic patient. *Journal of Experimental Psychology-Human Perception and Performance, 23*(4), 1217-1231. doi: Doi 10.1037//0096-1523.23.4.1217

Patterson, K., & Morton, J. (1985). From orthography to phonology: An attempt at an old interpretation. In K. Patterson, J. Marshall & M. Coltheart (Eds.), *Surface Dyslexia* (pp. 335-359). Hillsdale, N.J: Lawrence Erlbaum Associates.

Paulesu, E., Demonet, J. F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., . . . Frith, U. (2001). Dyslexia: cultural diversity and biological unity. *Science, 291*(5511), 2165-2167. doi: 10.1126/science.1057179

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S. F., . . . Frith, U. (2000). A cultural effect on brain function. *Nature Neuroscience, 3*(1), 91-96.

Peereman, R., & Content, A. (1998). Quantitative analyses of orthography to phonology mapping in English and French. 2014, from http://homepages.vub.ac.be/acontent/OPMapping.html

Perry, C., & Ziegler, J. (2002). Cross-language computational investigation of the length effect in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 28*(4), 990-1001. doi: 10.1037//0096-1523.28.4.990

Perry, C., Ziegler, J., Braun, M., & Zorzi, M. (2010). Rules versus statistics in reading aloud: New evidence on an old debate. *European Journal of Cognitive Psychology, 22*(5), 798-812.

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.

Perry, C., Ziegler, J., & Zorzi, M. (2014). When silent letters say more than a thousand words: An implementation and evaluation of CDP plus plus in French. *Journal of Memory and Language, 72*, 98-115. doi: 10.1016/J.Jml.2014.01.003
Plaut, D. C. (1999). A connectionist approach to word reading and acquired dyslexia: Extension to sequential processing. *Cognitive Science, 23*(4), 543-568. doi: 10.1016/S0364-0213(99)00015-4
Plaut, D. C. (2012). Giving theories of reading a sporting chance. *Behavioral and Brain Sciences, 35*(5), 301-302. doi: 10.1017/S0140525x12000301
Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychol Rev, 103*(1), 56-115. doi: 10.1037/0033-295x.103.1.56
Prinzmetal, W., Treiman, R., & Rho, S. H. (1986). How to See a Reading Unit. *Journal of Memory and Language, 25*(4), 461-475. doi: 10.1016/0749-596x(86)90038-0
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
Protopapas, A., Gerakaki, S., & Alexandri, S. (2006). Lexical and default stress assignment in reading Greek. *Journal of Research in Reading, 29*(4), 418-432. doi: 10.1111/J.1467-9817.2006.00316.X
Protopapas, A., & Vlahou, E. L. (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behavior Research Methods, 41*(4), 991-1008. doi: Doi 10.3758/Brm.41.4.991
R Core Team, R. (2013). R: A language environment for statistical computing [Computer software manual]. Vienna. Retrieved from http://www.R-project.org/
Rastle, K., & Coltheart, M. (1998). Whammies and double whammies: The effect of length on nonword reading. *Psychonomic Bulletin & Review, 5*(2), 277-282.
Rastle, K., & Coltheart, M. (1999). Serial and strategic effects in reading aloud. *Journal of Experimental Psychology-Human Perception and Performance, 25*(2), 482-503.
Rastle, K., & Coltheart, M. (2000). Lexical and nonlexical print-to-sound translation of disyllabic words and nonwords. *Journal of Memory and Language, 42*(3), 342-364. doi: 10.1006/Jmla.1999.2687
Rau, A., 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. doi: 10.1016/j.jcep.2014.09.012
Read, C., Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition, 24*, 31-44.
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
Rey, A., Ziegler, J., & Jacobs, A. (2000). Graphemes are perceptual reading units. *Cognition, 74*, 1-12.
Richlan, F. (2014). Functional neuroanatomy of developmetnal dyslexia: the role of orthographic depth. *Frontiers in Human Neuroscience, 8*, 1-13. doi: 10.3389/fnhum.2014.00347

Roberts, M., Rastle, K., Coltheart, M., & Besner, D. (2003). When parallel processing in visual word recognition is not enough: New evidence from naming. *Psychonomic Bulletin & Review, 10*(2), 405-414. doi: 10.3758/bf03196499

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

Rosenthal, R. (1979). The "File Drawer Problem" and Tolerance for Null Results. *Psychological Bulletin, 86*(3), 638-641.

Rosson, M. B. (1985). The Interaction of Pronunciation Rules and Lexical Representations in Reading Aloud. *Memory & Cognition, 13*(1), 90-99. doi: Doi 10.3758/Bf03198448

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: Doi 10.3758/Pbr.16.2.225

Schlapp, U., & Underwood, G. (1988). Reading, spelling, and two types of irregularity in word recognition. *Journal of Research in Reading, 11*, 120-132.

Schmalz, X., Marinus, E., & Castles, A. (2013). Phonological decoding or direct access? Regularity effects in lexical decisions of Grade 3 and 4 children. *Quarterly Journal of Experimental Psychology, 66*(2), 338-346. doi: 10.1080/17470218.2012.711843

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. doi: 10.1080/20445911.2014.968161

Seidenberg, M. (2011). Reading in different writing systems: One architecture, multiple solutions. In P. McCardle, B. Miller, J. Lee & O. Tzeng (Eds.), *Dyslexia across languages* (pp. 146-168). Baltimore: Paul Brookes Publishing.

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

Seidenberg, M., Waters, G., Barnes, M., & Tanenhaus, M. (1984). When Does Irregular Spelling or Pronunciation Influence Word Recognition? *Journal of Verbal Learning and Verbal Behavior, 23*(3), 383-404.

Seva, N., Monaghan, P., & Arciuli, J. (2009). Stressing what is important: Orthographic cues and lexical stress assignment. *Journal of Neurolinguistics, 22*(3), 237-249. doi: 10.1016/J.Jneuroling.2008.09.002

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

Share, D. (1995). Phonological recoding and self-teaching: *sine qua non* of reading acquisition. *Cognition, 55*, 151-218.

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.

Snowling, M. J. (2000). *Dyslexia*. Malden, MA: Blackwell Publishers.

Sprenger-Charolles, L., Siegel, L., Jiménez, J., & Ziegler, J. (2011). Prevalence and Reliability of Phonological, Surface, and Mixed Profiles in Dyslexia: A Review of Studies Conducted in Languages Varying in Orthographic Depth. *Scientific Studies of Reading, 15*(6), 498-521.

Sucena, A., Castro, S. L., & Seymour, P. (2009). Developmental dyslexia in an orthography of intermediate depth: the case of European Portuguese. *Reading and Writing, 22*(7), 791-810. doi: 10.1007/s11145-008-9156-4

Sulpizio, S., Arduino, L. S., Paizi, D., & Burani, C. (2013). Stress Assignment in Reading Italian Polysyllabic Pseudowords. *Journal of Experimental Psychology-Learning Memory and Cognition, 39*(1), 51-68. doi: 10.1037/A0028472

Tabossi, P., & Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Memory & Cognition, 20*(3), 303-313.

Taft, M. (1979). Lexical access via an orthographic code: The basic orthographic syllable structure (BOSS). *Journal of Verbal Learning and Verbal Behavior, 18*(1), 21-39.

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

Taft, M. (1992). The Body of the Boss - Subsyllabic Units in the Lexical Processing of Polysyllabic Words. *Journal of Experimental Psychology-Human Perception and Performance, 18*(4), 1004-1014. doi: 10.1037//0096-1523.18.4.1004

Taft, M., & Radeau, M. (1995). The Influence of the Phonological Characteristics of a Language on the Functional Units of Reading - a Study in French. *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale, 49*(3), 330-348. doi: 10.1037/1196-1961.49.3.330

Tan, L. H., & Perfetti, C. A. (1998). Phonological codes as early sources of constraint in Chinese word identification: A review of current discoveries and theoretical accounts. *Reading and Writing, 10*(3-5), 165-200. doi: 10.1023/A:1008086231343

Taylor, J. S. H., Plunkett, K., & Nation, K. (2011). The influence of consistency, frequency, and semantics on learning to read: An artificial orthography paradigm. *J Exp Psychol Learn, 37*(1), 60-76. doi: 10.1037/A0020126

Thaler, V., Ebner, E. M., Wimmer, H., & Landerl, K. (2004). Training reading fluency in dysfluent readers with high reading accuracy: Word specific effects but low transfer to untrained words. *Annals of Dyslexia, 54*(1), 89-113. doi: 10.1007/S11881-004-0005-0

Thompson, G. B., Connelly, V., Fletcher-Flinn, C. M., & Hodson, S. J. (2009). The nature of skilled adult reading varies with type of instruction in childhood. *Memory & Cognition, 37*(2), 223-234. doi: 10.3758/mc.37.2.223

Torgesen, J. K., Wagner, R., & Rashotte, C. (1999). TOWRE-2: Test of Word Reading Efficiency. Austin, TX: Pro-Ed.

Treiman, R., Goswami, U., & Bruck, M. (1990). Not All Nonwords Are Alike - Implications for Reading Development and Theory. *Memory & Cognition, 18*(6), 559-567. doi: Doi 10.3758/Bf03197098

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

Treiman, R., Kessler, B., & Evans, R. (2007). Anticipatory conditioning of spelling-to-sound translation. *Journal of Memory and Language, 56*(2), 229-245. doi: 10.1016/j.jml.2006.06.001

Treiman, R., Kessler, B., Zevin, J. D., Bick, S., & Davis, M. (2006). Influence of consonantal context on the reading of vowels: Evidence from children. *Journal of Experimental Child Psychology, 93*(1), 1-24. doi: Doi 10.1016/J.Jecp.2005.06.008

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. (1995). The Special Role of Rimes in the Description, Use, and Acquisition of English Orthography. *Journal of Experimental Psychology: General, 124*(2), 107-136.

Treiman, R., & Zukowski, A. (1991). Levels of Phonological Awareness. *Phonological Processes in Literacy*, 67-83.

Tressoldi, P. E., Vio, C., & Iozzino, R. (2007). Efficacy of an intervention to improve fluency in children with developmental dyslexia in a regular orthography. *Journal of Learning Disabilities, 40*(3), 203-209. doi: 10.1177/00222194070400030201

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.

Vaessen, A., Bertrand, D., Tóth, D., Csépe, V., Faísca, L., Reis, A., & Blomert, L. (2010). Cognitive development of fluent word reading does not qualitatively differ between transparent and opaque orthographies. *Journal of Educational Psychology, 102*(4), 827-842. doi: 10.1037/a0019465

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.

Venezky, R. L. (1970). *The structure of English orthography* (Vol. 82): Walter de Gruyter.

Wang, H. C., Castles, A., & Nickels, L. (2012). Word regularity affects orthographic learning. *Quarterly Journal of Experimental Psychology, 65*(5), 856-864. doi: 10.1080/17470218.2012.672996

Waters, G., & Seidenberg, M. (1985). Spelling-sound effects in reading: Time-course and decision criteria. *Memory & Cognition, 13*(6), 557-572.

Waters, G., Seidenberg, M., & Bruck, M. (1984). Children's and adults' use of spelling-sound information in three reading tasks. *Memory & Cognition, 12*(3), 293-305. doi: 10.3758/bf03197678

Weekes, B. (1997). Differential Effects of Number of Letters on Word and Nonword Naming Latency. *The Quarterly Journal of Experimental Psychology, 50A*(2), 439-456.

Wentink, H. W. M. J., VanBon, W. H. J., & Schreuder, R. (1997). Training of poor readers' phonological decoding skills: Evidence for syllable-bound processing. *Reading and Writing, 9*(3), 163-192. doi: 10.1023/A:1007921805360

Wimmer, H. (1993). Characteristics of Developmental Dyslexia in a Regular Writing System. *Applied Psycholinguistics, 14*(1), 1-33. doi: 10.1017/S0142716400010122

Wimmer, H. (1996). The Nonword Reading Deficit in Developmental Dyslexia: Evidence from Children Learning to Read German. *Journal of Experimental Child Psychology, 61*, 80-90.

Wimmer, H. (2006). Don't neglect fluency! *Developmental Science, 9*(5), 447-448.

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: word recognition in English and German children. *Cognition, 51*(1), 91-103.

Wimmer, H., Landerl, K., Linortner, R., & Hummer, P. (1991). The relationship of phonemic awareness to reading acquisition: More consequence than precondition but still important. *Cognition, 40*, 219-249.

Wimmer, H., Mayringer, H., & Landerl, K. (2000). The double-deficit hypothesis and difficulties in learning to read a regular orthography. *Journal of Educational Psychology, 92*(4), 668-680. doi: 10.1037//0022-0663.92.4.668

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual Differences in Visual Word Recognition: Insights From the English Lexicon Project. *Journal of Experimental Psychology-Human Perception and Performance, 38*(1), 53-79. doi: Doi 10.1037/A0024177

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

Ziegler, J., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., . . . Blomert, L. (2010). Orthographic Depth and Its Impact on Universal Predictors if Reading: A Cross-Language Investigation. *Psychological Science, 21*(4), 551-559. doi: 10.1177/0956797610363406

Ziegler, J., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3-29. doi: 10.1037/0033-2909.131.1.3

Ziegler, J., & Goswami, U. (2006). Becoming literate in different languages: similar problems, different solutions. *Developmental Science, 9*(5), 429-253.

Ziegler, J., Jacobs, A., & Stone, G. (1996). Statistical analysis of the bidirectional inconsistency of spelling and sound in French. *Behaviour Research Methods, Instuments & Computers, 28*(4), 504-515.

Ziegler, J., & Perry, C. (1998). No more problems in Coltheart's neighbourhood: resolving neighbourhood conflicts in the lexical decision task. *Cognition, 68*, B53-B62.

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.

Ziegler, J., Perry, C., Jacobs, A. M., & Braun, M. (2001). Identical Words are Read Differently in Different Languages. *Psychological Science, 12*(5), 379-384. doi: 10.1111/1467-9280.00370

Ziegler, J., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology, 86*, 169-193.

Ziegler, J., Perry, C., & Zorzi, M. (2014). Modelling reading development through phonological decoding and self-teaching: implications for dyslexia. *Philosophical Transactions of the Royal Society B-Biological Sciences, 369*(1634). doi: 10.1098/Rstb.2012.0397

Ziegler, J., 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. *Behavior Research Methods Instruments & Computers, 29*(4), 600-618. doi: 10.3758/bf03210615

Zoccolotti, P., De Luca, M., Di Pace, E., Judica, A., & Orlandi, M. (1999). Markers of developmental surface dyslexia in a language (Italian) with high grapheme-phoneme correspondence. *Applied Psycholinguistics, 20*(2), 191-216.

Zorzi, M. (2010). The connectionist dual process (CDP) approach to modelling reading aloud. *European Journal of Cognitive Psychology, 22*(5), 836-860. doi: 10.1080/09541440903435621