Universals in Learning to Read Across Languages and Writing Systems

Ludo Verhoeven & Charles Perfetti

To cite this article: Ludo Verhoeven & Charles Perfetti (2021): Universals in Learning to Read Across Languages and Writing Systems, Scientific Studies of Reading, DOI: 10.1080/10888438.2021.1938575

To link to this article: https://doi.org/10.1080/10888438.2021.1938575
Universals in Learning to Read Across Languages and Writing Systems

Ludo Verhoeven\textsuperscript{a,b} and Charles Perfetti\textsuperscript{a,b}

\textsuperscript{a}Behavioural Science Institute, Radboud University, Nijmegen, Netherlands; \textsuperscript{b}Learning Research & Development Center, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

ABSTRACT

In this article, we provide a cross-linguistic perspective on the universals and particulars in learning to read across seventeen different orthographies. Starting from the assumption that reading reflects a learned sensitivity to the systematic relationships between the surface forms of words and their meanings, we chose a broad group of seventeen languages, representing syllabic, morphosyllabic, alphasyllabic (abugida), abjad, and alphabetic writing systems. We investigated the systematic variation among these languages in their written forms and in their mapping of writing units to language units, and demonstrated the universality of operating principles in learning to read across languages and writing systems.

Reading engages cognitive language processes that are initiated by visual input from graphic strings. Cognitive and neurocognitive research has provided a variety of theoretical models for these processes and their acquisition. Until recently, however, research on languages other than English was too sparse to provide evidence of generality across languages and models of word identification based on English were not applicable to other languages, as pointed out by Share (2008). A cross-linguistic perspective with dual lenses to view the language-specific properties of reading and the universal properties common to reading in all languages and orthographies can provide a fuller view of reading and learning to read. In a previous attempt, McBride (2016) compared reading and writing development and impairment across a range of languages, scripts, and contexts while taking an ecological perspective. Moreover, Daniels and Share (2018) compared the consequences of writing system variation on reading and dyslexia.

Our aim was to find universals in learning to read through systematic analyses of a variety of languages, writing systems, and orthographies. To achieve this aim, we brought together experts with a variety of multidisciplinary backgrounds (including neurocognitive science, cognitive psychology, developmental cognitive science, psycholinguistics, and education) to consider reading and its development in a variety of written languages. Our claim was that reading reflects a learned sensitivity to the systematic relationships between the surface forms of words and their meanings. This can be seen as a universal aspect of reading. Because writing systems vary in how they represent the languages that they encode and languages vary in their phonological, morphological and semantic properties, we expected some specific effects to reveal themselves across writing systems and orthographies. To address the role of universals and particulars in learning to read, we chose a broad grouping of languages and writing systems: Asian (Chinese, Japanese, Korean, Kannada), West Semitic (Arabic, Hebrew), Romance (Italian, French, Spanish), Germanic (German, Dutch, English), Slavic (Czech, Slovak, Russian), Greek, Finnish, and Turkish. These languages also represent the major writing systems, namely the syllabic, morpho-syllabic, alphasyllabic, abjad, and alphabetic writing systems.

CONTACT Ludo Verhoeven (l.verhoeven@pwo.ru.nl) Behavioural Science Institute, Radboud University Nijmegen, Thomas van Aquinostraat 4, Nijmegen 6525 GD, Netherlands

© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.
The result was a volume reporting on learning to read across these seventeen languages (Verhoeven & Perfetti, 2017). Table 1 shows these languages and the five writing systems they represent. Our inclusion of seventeen languages is but a small sample of the world’s languages and certainly under-represents the languages read in South Asia and Sub-Saharan Africa. Nevertheless, the languages we sampled are spoken by over 3.7 billion people, more than half the earth’s population.

In what follows, we examine variation among these languages in their written forms and in their mapping of writing units to language units. Common to these two variations are the accommodations that written languages make to features of the spoken language, thus affecting both the mapping system and the forms of its graphs. We conclude with a discussion of the operating principles that reflect the universals across languages and writing systems.

**Variation in written and spoken languages**

First, we consider properties of the writing systems and their orthographies. The 17 written languages reviewed in Verhoeven and Perfetti (2017) include examples of each of the world’s five major writing systems, as indicated in Table 1. A fuller assessment of these 17 languages is contained in the concluding chapter in a table spanning 13 pages (Perfetti & Verhoeven, 2017). This assessment includes key features of the spoken language, the written language, and research results on learning to read.

With this background, we turn to two aspects of writing variation that are important to reading. The first is the perceptual demands of reading the written graphs. The second is the mapping between written and spoken language forms.

**Graphic efficiency: number and complexity of graphs**

The role of graphic complexity is not often emphasized, perhaps because in alphabetic writing the forms of letters are relatively simple, especially in Latin-based alphabets. However, even within alphabetic writing systems there is a range of complexity in the graphic forms and this range is dramatically larger when other systems are considered. From the perspective of the child learner, the graphs have specific properties that make them functional categories even for preliterate children. Children aged 3–5 understand that somehow an arrangement of graphs corresponds to a spoken word (Treiman, Hompleum, Gordon, Decker, & Markson, 2016) and differentiate the visual-spatial layout of their own writing system from others (Treiman, Mulqueeny, & Kessler, 2014). Thus graphs become

<table>
<thead>
<tr>
<th>Geographical Area</th>
<th>Language Group</th>
<th>Language</th>
<th>Writing system</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Asia</td>
<td>Sino-Tibetan</td>
<td>Chinese (Mandarin)</td>
<td>Morphosyllabic</td>
</tr>
<tr>
<td></td>
<td>Japanese</td>
<td>Japanese</td>
<td>Morphosyllabic (kanji)</td>
</tr>
<tr>
<td>South Asia</td>
<td>Koreanic</td>
<td>Korean</td>
<td>Syllabic (kana)</td>
</tr>
<tr>
<td>West Asia</td>
<td>Dravidian</td>
<td>Kannada</td>
<td>Alphabetic (syllabic alphabet)</td>
</tr>
<tr>
<td>East Africa</td>
<td>West Semitic</td>
<td>Arabic</td>
<td>Alphasyllabic</td>
</tr>
<tr>
<td>Europe and Intercontinental</td>
<td>Greek</td>
<td>Hebrew</td>
<td>Alphabetic: consonantal (abjad)</td>
</tr>
<tr>
<td></td>
<td>Latin</td>
<td>Italian</td>
<td>Alphabetic: consonantal (abjad)</td>
</tr>
<tr>
<td></td>
<td>Germanic</td>
<td>German</td>
<td>Alphabetic</td>
</tr>
<tr>
<td></td>
<td>Slavic</td>
<td>Czech &amp; Slovak</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Turkic</td>
<td>Finnish</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Turkish</td>
<td></td>
</tr>
</tbody>
</table>
perceptual categories that function as perceptual objects in learning to read. Being able to distinguish one from the other is fundamental. Thus, the visual complexity of the graphs is one of the challenges of learning to read and this challenge varies across orthographies.

The visual complexity of graphs can be measured by various operations on the forms of graphs. One useful objective measure is perimetric complexity, which captures the density of the lines of a graph (“black ink”) relative to the background space of the graph (Pelli, Burns, Farell, & Moore-Page, 2006). More formally, it is the ratio of the square of the sum of inside and outside perimeters to the product of $4\pi$ and the area of the foreground (Pelli et al., 2006; Watson, 2012). Perimetric complexity has been used to control graphic complexity in studies on learning to read (e.g., Liu, Chen, & Wang, 2016; Wang, McBride-Chang, & Chan, 2014; Yin & McBride, 2015). GraphCom is an aggregate measure developed to accommodate features of both non-alphabetic and alphabetic graphs that are not captured by perimetric complexity (Chang, Chen, & Perfetti, 2017). This measure, in addition to perimetric complexity, includes the number of simple features, number of connected points, and number of disconnected components. GraphCom is a single measure combining these four aspects of graphic form and was applied to 21,550 graphs from 131 written languages. It proved to be the best predictor of visual similarity judgments carried out on graphs sampled from languages representing the five major different writing systems (Chang et al., 2017).

The orthographies reviewed in Verhoeven and Perfetti (2017) show variation in graphic complexity across measures from Chang et al. (2017), as shown in Table 2. The table orders orthographies according to their complexity as measured by GraphCom. It also shows the four measures that make up GraphCom for each written language along with the GraphCom score, the average of the standardized scores of each of the four basic measures. Among the 17 languages, Hebrew has the lowest average graph complexity (GraphCom score). Although Arabic, like Hebrew, is an abjad consonantal orthography, its graphs are substantially more complex and are connected through

<table>
<thead>
<tr>
<th>Orthography</th>
<th>Graph inventory</th>
<th>Perimetric complexity</th>
<th>Number of disconnected components</th>
<th>Number of connected points</th>
<th>Number of simple features</th>
<th>GraphCom score and rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>6097</td>
<td>32.47</td>
<td>4.55</td>
<td>11.64</td>
<td>12.50</td>
<td>+5.59 1</td>
</tr>
<tr>
<td>Chinese</td>
<td>6097</td>
<td>29.47</td>
<td>4.01</td>
<td>9.54</td>
<td>10.60</td>
<td>+4.54 2</td>
</tr>
<tr>
<td>Japanese kanji</td>
<td>1006</td>
<td>28.62</td>
<td>3.84</td>
<td>9.65</td>
<td>10.43</td>
<td>+4.41 3</td>
</tr>
<tr>
<td>Japanese kana (hiragana)</td>
<td>48</td>
<td>23.32</td>
<td>1.29</td>
<td>2.75</td>
<td>4.19</td>
<td>+0.83 4</td>
</tr>
<tr>
<td>Kannada</td>
<td>50</td>
<td>12.55</td>
<td>1.42</td>
<td>2.40</td>
<td>3.84</td>
<td>+0.23 5</td>
</tr>
<tr>
<td>Korean</td>
<td>40</td>
<td>14.71</td>
<td>1.38</td>
<td>2.15</td>
<td>3.40</td>
<td>+0.21 6</td>
</tr>
<tr>
<td>Japanese kana (katakana)</td>
<td>48</td>
<td>16.06</td>
<td>1.38</td>
<td>1.56</td>
<td>2.96</td>
<td>+0.11 7</td>
</tr>
<tr>
<td>Arabic</td>
<td>28</td>
<td>8.78</td>
<td>1.82</td>
<td>1.36</td>
<td>3.07</td>
<td>−0.07 8</td>
</tr>
<tr>
<td>Russian</td>
<td>33</td>
<td>7.51</td>
<td>1.12</td>
<td>2.05</td>
<td>2.89</td>
<td>−0.43 9</td>
</tr>
<tr>
<td>Finnish</td>
<td>28 (32)</td>
<td>7.00 (+)</td>
<td>1.20 (+)</td>
<td>1.41</td>
<td>2.34 (+)</td>
<td>−0.55 10</td>
</tr>
<tr>
<td>French</td>
<td>26 (37)</td>
<td>6.85 (+)</td>
<td>1.04 (+)</td>
<td>1.44</td>
<td>2.23 (+)</td>
<td>−0.60 11</td>
</tr>
<tr>
<td>Spanish</td>
<td>27</td>
<td>6.93</td>
<td>1.07</td>
<td>1.48</td>
<td>2.31</td>
<td>−0.61 12</td>
</tr>
<tr>
<td>Greek</td>
<td>24</td>
<td>7.09</td>
<td>1.06</td>
<td>1.43</td>
<td>2.27</td>
<td>−0.62 13</td>
</tr>
<tr>
<td>German</td>
<td>26 (30)</td>
<td>6.85 (+)</td>
<td>1.04 (+)</td>
<td>1.44</td>
<td>2.25 (+)</td>
<td>−0.63 14</td>
</tr>
<tr>
<td>Dutch</td>
<td>26</td>
<td>6.85</td>
<td>1.04</td>
<td>1.44</td>
<td>2.25</td>
<td>−0.64 15.5</td>
</tr>
<tr>
<td>English</td>
<td>26</td>
<td>6.85</td>
<td>1.04</td>
<td>1.44</td>
<td>2.25</td>
<td>−0.64 15.5</td>
</tr>
<tr>
<td>Italian</td>
<td>21</td>
<td>6.74</td>
<td>1.02</td>
<td>1.45</td>
<td>2.17</td>
<td>−0.67 17</td>
</tr>
<tr>
<td>Hebrew</td>
<td>22</td>
<td>4.90</td>
<td>1.07</td>
<td>0.89</td>
<td>1.96</td>
<td>−0.87 18</td>
</tr>
</tbody>
</table>

(1) Data for Czech, Slovak and Turkish are not available. Czech (42 letters) and Slovak (46 letters) would rank as two of the most complex of the European alphabets in this sample. (2) Alphabet graph inventories show the number of basic letter shapes excluding diacritics. For those alphabets whose diacritics are not included in the graph inventory number, the total count that includes diacritic graphs is in parentheses and is included in estimating the GraphCom rank. (3) GraphCom is a derived score, the average for each language of the standardized scores for each of four other measures (Perimetric complexity, number of disconnected components, number of connected points, and number of simple features).
ligature. Alphabet graphs are relatively simple for orthographies based on the Latin alphabet used in modern European writing, with variations associated with the use of diacritics in some languages (Spanish, French, German) but not others (Dutch, English). The Russian Cyrillic alphabet has substantially more average complexity than the Latin based alphabets and the Korean Hangul alphabet has the highest complexity for an alphabet. Chinese has – by far – the highest average graph complexity as measured by every one of the four basic measures (thus GraphCom as well). Although the simplified character system used in the Chinese mainland reduces the GraphCom score of the traditional system substantially, the simplified Chinese is still much more complex than Kannada, the next most complex. Thus, Chinese is an outlier orthography in the average complexity of its graphs (characters) according to all the component measures as well as the overall GraphCom measure.

A comparison of Korean and Kannada is informative concerning the relationship of Perimetric Complexity and GraphCom. According to Perimetric Complexity, alphabetic Korean graphs are slightly more complex (14.71), on average, than graphs in Kannada (12.55), an alphasyllabary. However, given the fact that GraphCom includes the number of disconnected components, number of connected points, and simple features, it reverses this order, making Kannada graphs slightly higher on average.

We should expect graph complexity to affect the acquisition of high-quality orthographic representations (Perfetti, 2007), thus contributing to difficulty in learning to read. Especially important is that orthographies with visually complex graphemes contain a larger set of graphs, providing an additional source of difficulty during learning. The assumption that high levels of graphic complexity add to the challenges of learning to read can be quantified at a coarse grain. The 20–45 graphs typical in alphabets are mastered early in schooling, either as symbols with names or as symbols with pronunciation or both, depending on literacy instruction. Although learning the more complex 46 graphs of Slovak may present a greater challenge than the 26 less complex graphs of English, this challenge is met within similar approaches to early literacy that introduce children to the letters. The situation is different in alphasyllabaries, where several years of instruction are needed to introduce the much larger number of graphs (Nag, 2007, 2014) and thus affect the pace of learning to read (Nag, Caravolas, & Snowling, 2011).

Visual complexity is not something arbitrarily associated with writing systems. To paraphrase the idea that languages get the writing systems they deserve, writing systems get the complexity they need. The more graphs a writing system requires, the more complex the average graph must be. This is because each graph must be discriminable from every other graph and this discriminability can be achieved only by the accrual of many distinctive features. Indeed, the correlation between number of graphs and mean graphic complexity is \( r = .79 \) across the 131 written languages reported in Chang et al. (2017) and is \( r = .83 \) for our smaller sample here. The final link is the connection between the number of graphs and the mapping system. The principles of the mapping system largely determine the number of graphs required. On the language side of the mapping function, the number of language units is ordered from lowest to highest level: phoneme, syllable, morpheme, word. Thus, because they map phonemes, alphabets require the fewest graphs whereas morphosyllabaries, because they map syllable-size morphemes, require the most.

**The mapping principles of written language**

Writing encodes (spoken) language. The mapping principles by which writing encodes language are defined by the linguistic levels at which the written units connect: A morphemic or logographic principle connects writing units to meaning-bearing morphemes or multi-morpheme words; the syllabic principle, the highest level of speech-based mapping, connects the writing unit to a spoken syllable; the alphabetic principle connects the writing unit to the phoneme. However, writing also mixes these principles, combining phonemic and syllabic units (alpha-syllabaries) and making morphemes prominent across phoneme units (abjads) and syllables (morphosyllabaries). Writing systems
are embedded in cultural traditions and resist change, except in cases of authority-directed change, as is the case in two of our languages: Korean, which largely exchanged its Chinese morphosyllabary for an alphabet in the 15th century, and Turkish, which traded its Arabic based system for a Latin alphabet in the early 20th century.

Because writing is encoded language, not simply encoded speech, it can accommodate multiple aspects of a linguistic system, not just its phoneme inventory. Indeed, the specific form of writing sometimes suggests accommodations to features of language, managing to balance higher-level language units (morphology) with phonology (Frost, 2012; Perfetti & Harris, 2013; Seidenberg, 2011). The linguistic factors of a language, such as its phoneme inventory, syllable inventory, and morphological complexity appear to be accommodated by corresponding aspects of its written languages (Perfetti & Harris, 2013). A particularly strong hypothesis by Frost (2012) is that each language has its optimal writing system. However, because of the many influences on writing systems – cultural, religious, commercial, political – linguistic optimization is unlikely. Instead, the result may be better characterized as a writing system that accommodates language features sufficiently to (usually) resist radical changes.

Table 3 offers suggestive accommodations of our 17 written languages to their spoken languages. For example, Chinese may be well suited for its morphosyllabic writing system because of its many single-syllable morphemes, which create multiple meaning mappings for any given syllable. Alphabetic writing would create many homophones with identical spelling, whereas the character system identifies a particular morpheme by a (usually) unambiguous orthography.¹

Although one might think that English would be easier to read with a syllabary, its phonological complexity would create a large number of syllables and thus be much less efficient than an alphabet. Japanese, however, seems well suited for its kana syllabaries because many (but not all) of its syllables have simple consonant-vowel structure. Its use of morphosyllabic Chinese characters as its primary writing system may have its own advantages, although this is less clear, partly because reading characters can be based on their mapping to Chinese syllables or on their mapping to Japanese words. Written Hebrew and Arabic, as argued by Frost (2012), show an adaptation to the nonlinear consonant root structure of their spoken languages, allowing the letters of the root morpheme to become salient across multiple words that are derived from the root. Kannada uses an alphasyllabary which accommodates syllable variety. Finnish uses a transparent alphabet to map its many vowels and to write long words expressing complex inflectional morphology. The alphabets of Romance, German and Slavic languages vary in transparency and syllable complexity due to their response to preservation of morphology and to historical assimilation processes.

Operating principles in learning to read

Based on our analysis of these seventeen languages and their writing systems, we concluded that high-level empirical generalization applied to learning to read in all those languages and suggestively to reading in all languages (Perfetti and Verhoeven [2017]. The basic assumption is that reading development implies learning how a writing system encodes language. The general operating principles are what supports learning this encoding, enabling children to perceive, analyze, and use written language in ways that lead to the mastery of a particular orthography. We suggest these operating principles enable the processing of linguistic input and the organization and re-organization of representations that allow the learner to acquire knowledge of the writing system – spoken language connections for any language. Operating principles apply to the three major aspects of learning to read: becoming linguistically aware, acquiring word identification skill and learning to comprehend.

Operating principles for becoming linguistically aware

To become linguistically aware, children need to attend to the sounds of spoken language in order to uncover salient syllabic and phonemic boundaries in words and to search for written language signals for their connection to language.
Further consistent, written OP1; Medium Important Canonical Modest Phonologically Large Writing Large Table Complex

<table>
<thead>
<tr>
<th>Orthography</th>
<th>Important Language features</th>
<th>Writing System Accommodation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>Tonal language with simple syllabic structure. Syllables are morphemes, creating many homophones. Productive compounding.</td>
<td>Alphabets and syllabaries would have many homophones. Characters can distinguish between them.</td>
</tr>
<tr>
<td>Korean</td>
<td>Salient syllable body (CV nucleus). Simple consonants. Rich compounding and inflectional morphology. Most morphemes are monosyllabic.</td>
<td>With square syllable arrangement, alphabet accommodates better than original Chinese characters</td>
</tr>
<tr>
<td>Kannada</td>
<td>Canonical syllabic structure with simple and complex syllables. Agglutinating grammar with rich derivation and inflectional morphology.</td>
<td>Alphasyllabary. Akshara set accommodates syllable variety. Writing encodes morphophonology.</td>
</tr>
<tr>
<td>Arabic</td>
<td>Simple syllabic structure with CV and CVC as basic types. Rich derivational and inflectional morphology in a nonlinear consonant root structure.</td>
<td>Arabic abjad displays the consonant morpheme roots along with optional diacritics for vowels and grammatical endings.</td>
</tr>
<tr>
<td>Hebrew</td>
<td>Phonological complexity, multisyllabic with extensive consonant clusters. Rich derivational morphology, including productive nonlinear consonant root system (similar to Arabic).</td>
<td>Abjad orthography allows salient expression of the morpheme root consonants with optional vowel representations.</td>
</tr>
<tr>
<td>Greek</td>
<td>Complex syllable structure in mostly multisyllabic words. Highly productive derivation and compounding, and rich inflectional morphology.</td>
<td>Highly consistent alphabetic orthography with minor complexities. Some phonemes map to more than one graph.</td>
</tr>
<tr>
<td>Italian</td>
<td>Mainly open syllables with some consonant clusters. Productive word formation and rich inflectional morphology.</td>
<td>Highly consistent alphabetic orthography with morphemes preserved.</td>
</tr>
<tr>
<td>French</td>
<td>Large number of vowels. Mainly simple open syllables. Rich inflectional morphology often not phonologically expressed at word endings.</td>
<td>Alphabetic orthography has reduced transparency. Spellings carry morphological information not present in speech.</td>
</tr>
<tr>
<td>Spanish</td>
<td>Phonologically simple with open syllables. Inflectional morphology (typical of Romance languages).</td>
<td>Transparent orthography. Consistent mappings except for three graphemes.</td>
</tr>
<tr>
<td>German</td>
<td>Medium phonological complexity with mostly closed syllables that can be complex. Rich derivation and inflectional morphology.</td>
<td>Transparent orthography with preservation of morphological structures.</td>
</tr>
<tr>
<td>Dutch</td>
<td>Medium phonological complexity with mostly closed syllables that can be complex. Simple derivation and inflectional morphology.</td>
<td>Transparent orthography. Some loss of morpheme transparency in spellings that preserve pronunciation.</td>
</tr>
<tr>
<td>English</td>
<td>A large number of syllables and high phonological complexity. Simple inflectional morphology and morphophonemes favor morpheme spellings.</td>
<td>An alphabet that encodes phonology inconsistently and partly preserves morphology. Too many syllables for a syllabary.</td>
</tr>
<tr>
<td>Czech/Slovak</td>
<td>Two closely related mutually intelligible languages. Polysyllabic word structures with mostly open syllables but complex onsets. Rich word formation and inflectional morphology.</td>
<td>A largely transparent alphabetic orthography with some inconsistencies from the preservation of morphology.</td>
</tr>
<tr>
<td>Russian</td>
<td>Large consonant inventory but only 5 vowels. Mostly open syllables, but with complex consonant onsets. Very rich inflectional morphology.</td>
<td>Cyrillic alphabet with transparent orthography. Vowel letters in two forms to show pronunciation of the consonant.</td>
</tr>
<tr>
<td>Finnish</td>
<td>Strong syllabic structure with use of vowel harmony. Many diphthongs plus 8 simple vowels. Vowel and consonant lengths are phonemic. Complex derivational and inflectional morphology creates long words.</td>
<td>Highly transparent alphabet uniquely maps all the many vowel sounds and adapts to complex inflectional morphology.</td>
</tr>
<tr>
<td>Turkish</td>
<td>Modest phoneme inventory. Strong open syllabic structure with vowel harmony. Regular derivation and inflectional morphology.</td>
<td>Highly Transparent. Latin alphabet provides consistent mappings.</td>
</tr>
</tbody>
</table>

**OP1: attend to salient stretches of speech as indicated by stress, intonation, rhythm**

Learning to read requires a sensitivity to the units of spoken language so that they can be mapped by written units. The first operating principle is that attention to the sounds of human speech supports this sensitivity. With exposure to speech, infants learn to parse the incoming acoustic signal into consistent, replicable chunks that come to represent phonemes (cf. Kuhl et al., 1997; Poeppel, 2014). Further attending to salient stretches of speech, marked by stress, intonation, and rhythm allows the
development of high-quality speech-based lexical representations (see Bailey, Plunkett, & Scarpa, 1999). There is evidence that both the quality and quantity of the lexical representations develop with reading and support it (cf. Verhoeven & Perfetti, 2011). When the child is ready for literacy, sensitivity to speech units support picking out the units to be mapped to the graphs of the writing system. Sensitivity to salient stretches of speech has proved to be important not only for alphabetic reading but for syllabic and morphosyllabic reading as well (see also Goswami, 2000).

**OP2: attend to any salient syllabic, onset-rime, or phoneme boundaries in words**
The richness and specificity of developing phonological representations and the need to distinguish among them have a strong impact on awareness of syllabic, onset-rime, or phoneme boundaries in words and thus on lexical quality. Across languages we found that learning to read builds upon a child’s phonological awareness, i.e., their ability to attend to the sounds of language independent of meaning. Broadly, this awareness entails the ability to isolate words in sentences, but more narrowly, the ability to identify sublexical units – syllables, rhymes, the beginnings of words, the ends of words, and phonemes. Awareness of syllables emerges earlier than awareness of phonemes because of the greater acoustic salience of syllables (Treiman & Zukowski, 1996). It is interesting that this turns out to be true across languages that have very different syllable structures. The predictive power of syllable-level awareness across different writing systems and languages – including alphabets, which do not encode syllables – highlights the general importance of attention to the sound structure of a language for learning to read. The predictive power of phoneme-level awareness differs across writing systems. Across alphabetic orthographies, phonemic awareness appears to be uniformly important as predictor of learning to read (see Moll et al., 2014). Evidence in alphabetic languages for the late association between phonemic awareness and literacy suggests that phonemic awareness and learning to read alphabetically can develop reciprocally. This means that phonemic awareness is an enabler rather than a prerequisite for alphabetic reading. However, in reading syllabaries and morphosyllabaries phonemic awareness is not a uniformly important factor. In Japanese, associations between phoneme-level awareness and reading occur late whereas in Chinese they show variation, partly due to the varieties of spoken language and variation in instructional practice, especially the use of an alphabet (Pinyin) at the very beginning of reading. In summary, we found that attention to the sounds of a language, as independent of meaning, facilitates the distinct representation of multiple levels of speech units that can then provide the basis for the mapping functions of varying orthographies. This universal principle is in line with previous cross-linguistic research (see Caravolas, Lervåg, Defior, Seidlová Málková, & Hulme, 2013; McBride-Chang et al., 2008; Melby-Lervåg, Lyster, & Hulme, 2012; Moll et al., 2014; Perfetti & Harris, 2013; Landerl, Castles, & Parrila, this volume).

**OP3: attend to written language signals for their connection to language**
Across different languages and cultures, interaction with symbols in the environment and literate others helps children learn that print carries meaning, that written texts may have various forms and functions, and that ideas can be expressed with spontaneous (non)conventional writing. If learning to read entails learning how one’s writing system encodes one’s language, then children must learn at least the following: (1) the inventory of graphic forms for a given writing system, (2) the orthographic units (graphemes) that connect to spoken language units (phonemes and morphemes), and (3) how specific orthographic units map onto specific phonological and morphological units of the spoken language (cf. Ehri, 2014). At a very young age, children may thus discover written language as a new modality of communication. By attending to the print environment, children gain partial knowledge of these writing features prior to literacy instruction. Additional attention to the constituent sounds (phonemes, syllables, morphemes) and graphemes of familiar words can support learning of specific graph-to-language mappings and may also lead to generalizations about how their written forms map to their language. We concluded that such self-learning is applicable across writing systems and orthographies. In the case of morphosyllabic writing, children learn that word meanings consist of characters that are linked to smaller meaning units (morphemes). In the case of alphabetic writing,
children learn that words consist of letters that are linked to phonemes. And there is general agreement that the development of literacy and mastery of a writing system involves discovery of the principles of linguistic recoding (see also Perfetti, Liu, & Tan, 2005). Prior to instruction, children may begin with a limited collection of written words that have personal meaning. On the basis of the constituent sounds and graphs encountered in these words, children may discover the mapping principle of their writing system, even for an alphabetic system. And phonological recoding can be viewed as an inductive learning mechanism that helps children associate letters with sounds when they attempt to read a word. However, for alphabetic orthographies, research suggests that such self-discovery is of limited use for most children. A more systematic approach that directly instructs children on the alphabetic principle and specific grapheme-phoneme correspondences is more effective, as shown in many reviews (e.g., Castles, Rastle, & Nation, 2018).

**Operating principles for the development of word identification**

Beyond discovering the mapping principle of their writing system, children acquire precise connections of their orthographic inventory with language units, allowing them to compute orthographic representations and to gain orthographic fluency through reading.

**OP4: increase the orthographic inventory**

Depending on the writing system, learning to identify words may need an extended period for learning graphic forms. The importance of orthographic knowledge, sometimes neglected for alphabetic writing, spans from initial learning to later automatized word identification. The more graphs an orthography has, the more demanding orthographic learning will be. Chinese literacy instruction targets the 6000 graphs commonly in use among the simplified set and many more in dictionaries containing traditional characters. The massive grapheme inventory of Chinese requires learning of Chinese characters to continue for several years. Note that there is no clear analogy between the phonological and semantic radicals found in Chinese characters and words spelled using an alphabet in other languages. Chinese words rely more heavily on morpheme compounding, allowing semantic cues to those words with transparent morphemes (cf. McBride & Wang, 2015). In Japanese, orthographic learning is demanding because it involves both morpheme-based kanji and two forms of kana syllabaries in which 110 syllabic units (morae) are represented by 105 symbols. The orthographic demands are also high for alphasyllabaries, whose graphic forms can take several years to fully master, because of a generally high number of graphs and variation among consonant graphs containing implicit vowels. Alphabetic writing systems have the advantage of calling upon a relatively small inventory of graphs (letters) that can be mastered in the first year of instruction or even prior. During initial phonics instruction, children are given the opportunity to consolidate the grapheme-phoneme correspondence rules that they have acquired. An important cross-linguistic finding is that spelling practice helps children internalize orthographic structures. For Chinese, it has been shown that the repeated writing of characters can improve the quality of the underlying orthographic representations thus giving needed support to reading. And for alphabetic orthographies, spelling practice strengthens word decoding, even though spelling develops slower than reading (Guan, Liu, Chan, Ye, & Perfetti, 2011).

**OP5: increase the inventory of familiar words through reading**

In all languages, written words can become familiar perceptual objects that are then recognized more quickly. Learning to read fluently builds on this increasing familiarity (see Seidenberg, 2017). A mechanism that serves reading of both deeper and shallower orthographies is the self-teaching procedure identified by Share (2004). An encounter with a word leads to phonological recoding, which is then fed back to the word’s orthography, initiating the word-specific identification process. With this mechanism, only a few exposures to the same word can be sufficient for storage of the word’s orthographic representation. Turning the unfamiliar into the familiar is relatively simple in shallow alphabetic writing. The first encounter with a new written word leads to decoding of the written form into its phonological form, producing immediate familiarity; greater exposure may be needed to gain
familiarity with deeper orthographies. The result is word identification driven, in part, by a familiarity-based memory retrieval process that is universal. And if such familiarity-based retrieval is functional in syllabic or alphabetic reading, then its importance in a language such as Chinese may be even greater. Japanese children appear to quickly learn the Japanese syllabary, demonstrating that transparency at both syllable and phoneme levels supports learning. Building orthographic representations in Chinese and kanji is laborious, because single characters need to be memorized one by one (cf. Chen, Wang, & Luo, 2013), although limited transfer across related characters is possible. Among alphabetic orthographies, the speed and accuracy of reading familiar words is highly affected by both syllabic complexity and orthographic depth. Finnish, Turkish, Korean, Czech, Slovak, Greek, Spanish, Italian, Russian, Dutch and German all show rapid learning. French, Arabic and Hebrew seem to be slower, whereas English and Kannada turn out to be among the hardest to learn. These observations are commensurate with cross-linguistic research comparing the development of decoding skills (e.g., Aro & Wimmer, 2003; Seymour, Aro, & Erskine, 2003).

**OP6: to gain word identification fluency**

Across languages, word identification shifts from computation to memory-based retrieval for words when they become familiar. Generally speaking, reading speed becomes the distinguishing marker of skill once children have reached high accuracy for word identification and decoding. Both sublexical letter strings and whole-word retrieval cues are used by familiarity-based processes, which allow a rapid look-up procedure rather than a computational procedure, no matter the language or writing system. Beyond establishing words as familiar, reading experience can produce gains in word reading fluency (cf. Caravolas et al., 2013; Vaessen et al., 2010). Highly fluent word reading is an effortless perceptual response that can include the automatization of word decoding, familiarity-based memory retrieval, and the attainment of fluent skilled reading. And these developments allow cognitive resources to be redirected to comprehension.

Across different orthographies, the data show that parallel developmental gains in both word decoding speed and accuracy occur very rapidly after the start of explicit reading instruction, while steady improvements in the speed and accuracy of word decoding occur in the years thereafter (see Georgiou et al., 2020). Across languages, we find accuracy to be mainly predicted from phonological awareness, and speed from rapid automatized naming (see also Landerl et al., 2019; Moll et al., 2014). The initially rapid building of the speed and accuracy of word decoding together with their continued development thereafter is commensurate with the restricted interactive model of lexical development (Perfetti, 1992), in which word decoding increases the number of orthographically addressable words and modifies already developing word representations. The sharp increase in the speed and accuracy of word decoding during the early stages of reading instruction suggests, moreover, that specific word representations develop fast but further improvement entailing increased numbers of redundant phonemic representations occurs more slowly. Thus, the development of word decoding can be defined as the acquisition of increased specificity and redundancy for lexical representations.

**The development of reading comprehension**

The operating principles for the development of word identification apply to the foundational and thus restricted component of skilled reading. Skilled reading comprehension and its development are critical, but more complex and more challenging for creating a small set of principles. Although Verhoeven and Perfetti (2017) offer several such principles, here we restrict our focus to the importance of word meaning as the connection point from word identification to comprehension, as shown in the Reading Systems Framework (Perfetti & Stafura, 2014). Word meaning is a fuel for the word-by-word integrative processes that constitute the recurring basic process in comprehension and its development. We focus on three principles that are part of this linkage to comprehension and thus foundational. They concern morphological processes, the use of linguistic knowledge, and the retrieval of background knowledge.
**OP7: attend to morphological relations**

Morphological knowledge provides building blocks of meaning and meaning-related connections to grammatical systems. The most directly meaning based process comes from compounding, combining morphemes to create new meanings (e.g., “blackbird”) and is common across languages. Other derivational processes link meanings across grammatical form class (“a game,” “to game”). But even inflectional morphology is meaning-related. Morphological affixes, bound morphemes that are joined with a root or stem, provide meaning that is realized grammatically in number, tense, aspect, and other information. Morphological knowledge turns out to be variably associated with reading across languages and writing systems, consistent with the assumption that writing systems manage trade-offs between morphology and phonology in different ways (see Frost, 2012). Thus, reading morphosyllabic Chinese relies on knowledge of Chinese morphological meaning compounding. Reading Arabic relies on the tri-consonant root morphology expressed in its abjad system. Moreover, Perfetti and Verhoeven (2017) evidenced that even the most transparent alphabetic orthographies (Finnish, Turkish) show a morphological effect in learning to read, as evidenced by a high association between morphological awareness and reading comprehension. Despite their phonological transparency, Finnish and Turkish reading are associated with morphology because of their rich agglutinative morphology, which creates long multi-morpheme words. French reading and English reading are associated with morphology because their orthographies fail to provide consistent phonological mapping. French goes farther, conveying in writing some inflectional endings that are suppressed in the spoken language. Along with Dutch, more transparent than English but much less than Finnish, morphological knowledge adds to phonological awareness in enabling reading. A reader’s awareness of morphological decomposition arises from a learned sensitivity to the systematic relationships among the surface forms of words and their underlying meanings (cf. Seidenberg & Gonnerman, 2000).

Across languages, word identification is the selection of a unique word based on an orthographic form that is uniquely associated with a particular phonology, morphology, and meaning. However, the writing system and its specific orthography define the priorities of morphology and phonology and thus regulate their roles in learning to read. Across languages, morphological awareness was found to be associated with word-to-text integration. There is also cross-linguistic evidence that a morphological family, centered on a single-root morpheme, can have a role in identifying a word and integrating it with antecedents, by activation that spreads, coactivating morphological lines family members along with the specific word (cf. Rueckl, 2010).

**OP8: use knowledge of language as you read**

With increasing skill in word identification, reading is increasingly dependent on general knowledge of the language, most of which, at first, has come from spoken language. (This changes with increasing literacy.) Of course, knowledge of the spoken language, (oral language skill, as it is often called) is essential to get reading started. However, the more general concept of language knowledge continues to be important and may even increase in importance as texts become more demanding of both grammatical and vocabulary knowledge, both of which extend beyond what a child has prior to literacy.

The idea that reading comprehension is fully accounted by word identification and listening comprehension (Hoover & Gough, 1990) is part of this picture and applies across the languages we studied in Verhoeven and Perfetti (2017). Indeed, we found that across languages reading comprehension became more dependent on children’s oral language comprehension as their word decoding skills increased. However, this simple view is incomplete in accounting for development of reading skill because reading itself brings about the learning of vocabulary and experience with a wider variety of grammatical structures and text types that are not experienced in typical spoken language (outside of academic lectures). Reading also increases the general knowledge that is needed to support comprehension.
Thus, once word identification is established at a fluent level, language factors determine the cognitive contribution to reading comprehension; needed critically is sufficient engagement with what is being read to reach the potential of these factors. Orthographies can influence the relative importance of both language and general knowledge, the factors important for language comprehension. When a consistent orthography eases the demands of word reading so that accuracy is easily attained, reading comprehension becomes more dependent on these factors. Thus, just as writing systems manage trade-offs between morphology and phonology, related down-stream trade-off effects appear in comprehension.

The language factors include implicit linguistic knowledge about the language’s syntax and morphology to support the building and integrating of constituent phrases during sentence comprehension; these processes are critical in enabling the reader to obtain and integrate propositional meaning from texts (cf. Perfetti, 2007; Verhoeven & Perfetti, 2008). Knowledge of word meanings (vocabulary), which reflect joint contributions of linguistic and conceptual knowledge, is central in these comprehension processes. The importance of vocabulary for reading comprehension was evidenced for every language reported in Verhoeven and Perfetti (2017).

Because so much of vocabulary is acquired following beginning reading, it is not simply a store of language knowledge waiting to be unlocked by decoding. Word meanings are continuously being retrieved, learned, and fine-tuned by reading itself. Both the quality of specific word knowledge (lexical quality) and the quantity of known words are important in supporting comprehension (Perfetti, 2007; Perfetti & Hart, 2001). It might seem convenient to subsume vocabulary under spoken language comprehension and thereby have a two-factor model of reading comprehension. However, this would fail to capture some observations about word meanings. For example, vocabulary knowledge directly supports identification of words that have exceptional spellings (Ricketts, Nation, & Bishop, 2007). A model that allows a more direct influence of knowledge of word meanings on reading comprehension may be more appropriate across languages (see Verhoeven & van Leeuwe, 2008). Beyond beginning reading, where only spoken language vocabulary is available, word meanings are not intrinsically part of spoken language more than written language. In both cases, they are the central connection point between coded input and comprehension, as much a component of a reading system as a language system (Perfetti & Stafura, 2014).

In reading comprehension beyond the word level there are several complexities that remind the reader that word and language knowledge remain central even in these other areas. Across languages, we found that beginning readers focus on words and learn to attach each word in a sentence to a syntactic phrase and its referential meaning to a semantic representation in perspective of the construction of a text model (see also Cain & Barnes, 2017). The successful integration of meaning enables the reader to encode a linguistic representation of the text that can be enriched through the retrieval of background knowledge to generate a more specified mental model (see Oakhill, Cain, & McCarthy, 2015).

**OP9: mobilize executive skills to supplement the literal meaning of text with relevant background knowledge**

Because texts are never fully explicit, there are abundant occasions on which readers must make inferences about what a text is about and thereby draw upon their prior knowledge. It is this level of text comprehension that both requires and enables a situation model, calling for additional information to be incorporated into purely linguistic-based accounts of text comprehension. Across languages, we found that text titles, headings and paragraph structures may help to activate prior knowledge (see also Wiley & Rayner, 2000). The components of a situation model can include information from the text, inferences based on the text, relevant prior knowledge, and inferences connecting the text and prior knowledge. A situation model supports readers’ efforts to identify problems and generate problem-solving strategies, and observe the results of attempted solutions. A basic premise is that text comprehension across languages and writing systems requires a mental simulation of a referential situation, constrained by the information contained in a text, the processing capacity of the human brain, and the nature of human interaction (Li & Clariana, 2019). Across languages, we found that
executive functioning (see also Cutting, Materek, Cole, Levine, & Mahone, 2009; Nouwens, Groen, Kleemans, & Verhoeven, 2021) and working memory capacity (see also Cain, 2006; Nouwens, Groen, & Verhoeven, 2017), explain the individual variation in reading comprehension after taking word decoding and language skills variance into account. Importantly, it has also been found that situation model building ability predicts reading comprehension over and above the other cognitive and linguistic predictors (see Raudzus, Segers, & Verhoeven, 2019).

**Universals in learning to read**

Like many generalizations that aim to apply over wide variation, the operating principles we have suggested appear to be universal. Learning to read engages an emergent reading network in the brain that increasingly shows universal properties, including high overlap with brain areas that support spoken language (Paulesu et al., 2000; Rueckl et al., 2015). This convergence of speech and reading areas in the brain across unrelated languages reflects the broad universal principles that writing encodes language and that reading engages phonology. Both visual cortex and temporal/parietal networks for lexical and unification processing are recruited (see Dehaene & Cohen, 2011; Hagoort, 2017). Thanks to the (re)structuring (OP1) and increasing awareness (OP2) of the phonological infrastructure of spoken language, and as a result of a learned specialization to recognize (OP3) and extend orthographic codes (OP4), visual word forms are stored in memory which increase in number, specificity (OP5) and redundancy (OP6) through reading exposure. A connection can then be made between these lexical building blocks and basic unification blocks with morphological (OP7) and syntactic relations (OP8), on the one hand, and general knowledge (OP9), on the other. Given that unification principles are largely shared between reading and speech, they will become easily available to the reader in case automated mappings between orthography and phonology are established. Pending typical language development and adequate levels of executive functioning, this may lead to successful reading comprehension (see also Cain & Barnes, 2017).

It is interesting to note that the same operating principles have also been found to be relevant for reading disabilities. In an attempt to search for evidence of universals in developmental dyslexia across languages and writing systems, Verhoeven, Perfetti, and Pugh (2019) compared a subset of eleven out of the seventeen languages examined in this article and concluded that phonological problems in reading and spelling words (i.e., phonological problems underlying the OPs 1 to 6) can be considered a core deficit in developmental dyslexia. And a recent meta-analysis of studies examining the comprehension problems of children with poor comprehension despite adequate decoding by Spencer and Wagner (2018) showed that these children had a delay in oral language development (i.e., OPs 7–8). Claims for empirical universals in reading and reading problems do not imply that there are no specific influences of writing system, orthography, language, and knowledge. It is important to understand the particular ways that language and writing affect reading as well as the broad universal principles we have described here. As a case in point, there appear to be variations in parts of the network that reflect writing system factors (e.g., Perfetti, Cao, & Booth, 2013).

There is no doubt that writing and language factors affect learning to read. The incremental learning of many Chinese characters over extended years contrasts with the rapid learning of a small set of alphabetic letters that allows most words written in a transparent alphabet to be read within a few weeks of literacy instruction. On the other hand, the effect of reading experience (practice) is to allow reading in divergent systems to become increasingly fluent, based on retrieval of word identities from orthographic input. The story of learning to read thus is one of universals and particulars: (i) Universals, because writing maps onto language, no matter the details of the system, creating a common challenge in learning that mapping, and because experience leads to familiarity-based identification across languages. (ii) Particulars, because it does matter for learning how different levels of language – morphemes, syllables, phonemes – are engaged; this in turn depends on the structure of the language and how its written form accommodates this structure.
Note
1. This analysis simplifies a complex history of language change in Chinese that included its own accommodation to single-syllable homophony by becoming (long after the development of the character system) predominately a language of two-syllable words. Language change and other factors challenge strong claims that a present day writing system is an accommodation to present-day language. Nor do we claim that a specific writing solution is optimal for a language, given other writing devices. We emphasize that our analysis merely suggests that some languages have written forms that accommodate their linguistic properties.

References


