How children learn to use a writing system: Mapping evidence from an Indic orthography to written language in children's books

Sonali Nag

University of Oxford

To appear in Written Language and Literacy early-2022

Acknowledgment: This study was funded by a UKRI Collective Fund award to the author and project partners of the UKRI GCRF Supporting Oral Language Development Project (ES/T004118/1). I am grateful for the careful attention to detail in the development of the Promise Print Corpus by Research Associates Sanjana Nagendra, Megha Nakshathri, Kala B., Rayan Miranda, Robert D'Souza, Pooja Pandith, Deekshitha Kotian and The Promise Foundation, India; Research Affiliates Drs. Shivani Tiwari and Sunila John; Data Scientist Dr. Aakash Agarwal, and Prof. R. Amritavalli. An earlier version of this paper titled 'Emergent and early literacy: how children learn to use a writing system' was presented as a Key Note in the annual conference of the Association for Written Language and Literacy held at the Department of Classics, University of Cambridge in 2019.

Address for correspondence:

Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, UK.

sonali.nag@education.ox.ac.uk

How children learn to use a writing system: Mapping evidence from an Indic orthography to written language in children's books

Abstract

Decoding a writing system is an impressive task requiring recognition of connections between printed symbols and the language they represent. Recognising the linguistic anchors for individual symbols is however not enough. Inferences are needed about unseen and often unstated encoding principles. This paper reviews task demands implicit in children's books and finds that current models of orthographic learning in an Indic writing system must go beyond a focus on intra-symbol cues, the size of the symbol set, and the nature of sound-symbol mapping. The child-directed print corpus shows a substantial demand for visual word recognition of multimorphemic words. Since children encounter an ever-expanding variety of such words in the books they read, it is essential to mount systematic studies. At a methodological level, this exploratory study shows the limitations of building models of literacy development when real world encounters with a writing system are not adequately taken into account.

Key words: writing system, orthographic breadth, orthographic depth, emergent literacy, early literacy, akshara, child-directed print, children's books, inflectional morphology, intra-symbol processing

Books for children are a valuable source of information at several levels. Books at home, for example, can be a sensitive index of opportunities for children to practice reading and see the purposes of literacy for different family members. These affordances perhaps explain why the count of books at home is a robust predictor of individual differences in emergent, early and later literacy in multiple languages and writing systems (Nag, Vagh, Dulay & Snowling, 2018b; Nag, Vagh, Dulay, Snowling, Donolato, et al., in preparation). At another level, the content in children's books can reveal assumptions about children's background knowledge and the elements that draw children's attention to event sequences, character actions and the internal world of who-is-thinking-what-and-why. For example, a largely single-study evidence-base shows an association between types of illustration and children's narrative expression about a book story (Coates, 2017; Pearce, 2003; Shapiro & Hudson, 1991). Yet another level of analysis of children's books is related to the quality of language. The vocabulary and syntactic structures in books could reflect a tension between use of authentic language and the authors' estimations of children's language development. This tension between authentic language and language at the level of children's understanding is an understudied area, but the available evidence suggests that book language is more varied and complex than child-directed spoken language (Hayes & Ahrens, 1988; Massaro, 2015; Montag, 2019; Montag, et al., 2015). All these levels of book analysis—books as a resource, content in books, and book language—are also important to the study of children's skills with the writing system used in a book, although the information they can provide is only indirect. Instead, the analytic level best suited to examine the question of writing system usage is at the level of the individual words in books; this is because visual word recognition requires direct deciphering of a writing system. A word-level analysis can also reveal writing system exposure, and this information could be used to interrogate the extant evidence-base about emergent and early literacy.

In this paper, words in children's books are examined against what is known about the development of emergent and early literacy in Kannada, a Dravidian language spoken in south India and a widely spread-out diaspora. Kannada orthography belongs to the group of south and south-east Asian orthographies that use the Indic writing system. This writing system is an alphasyllabary (Bright, 1992), and is also called an abugida (Daniels, 1990) and the akshara writing system (Nag, 2017). Kannada is an agglutinative language with rich inflectional and derivational morphology. Understanding words in Kannada children's books is useful because, as with many other aspects within child psychology and applied linguistics, the field of literacy development remains undersupplied with studies from non-WEIRD[1] contexts. A study of demands on children's Kannada emergent and early literacy from the perspective of word encounters is also useful because the orthographies of several south and south-east Asian languages have a similar architecture to the Kannada orthography, and hence inferences in Kannada about an important contextual factor (the books children read) can be extended to a good selection of other languages. In addition, a review of written words in Kannada is useful because the orthography is different in important ways from others such as the European orthographies and Chinese; these contrasts allow orthography-specific and potentially universal inferences about what children's books offer for literacy development.

1. A Sketch of an Indic Orthography

Learning to read an Indic orthography, as with learning any writing system, draws upon at least two knowledge-bases: orthographic knowledge and linguistic knowledge. A knowledge-base about the orthographic system focuses on the symbol repertoire and the set of mapping principles that connect print to language. The knowledge-base about the

[1] WEIRD is the acronym for western, educated, industrialized, rich and democratic societies. For several decades, research from this narrow set of contexts has supplied the evidence-base in multiple areas of human development, prompting a call to 'structurally re-organise' the behavioural sciences (Henrich, Heine & Norenzayan, 2010).

linguistic system includes details about phonological representations, lexical identities of words, chunked linguistic information beyond words, and a meta-linguistic skill about how language works. An example of the mapping across orthography, phonology and basic semantics is seen, for example, in decoding the monosyllablic word ‘hen’. The phonology needs to be activated for each individual orthographic unit (<h>, <e>, <n>), blended to form the word, and links are to be made with the mental lexicon (after the ‘multicomponent processing system’ for computing meaning of words, in Harm & Seidenberg, 2004, p. 716). The Kannada word for hen, /koLi/, when written in the Kannada orthography, needs recognition of the orthographic unit that represents /ko/ separate from the orthographic unit for /Li/.[2] In addition, both <ko> and <Li> are symbol blocks because it is possible for each to be further broken up into <k> and <o>, and <L> and <i>. Knowing where to break up symbol strings in words into individual symbol blocks and each block into component markers is a necessary skill because representation in this writing system transcribes all the phonemes of a word but is orthographically organised in blocks. This form of orthographic representation is found in all Indic scripts of south and south-east Asia including Thai, Punjabi, Hindi, Bengali, Sinhala and Tamil. It is the symbol block that is called *akshara*.



Figure 1

Five example words in Brahmi constructed from akshara on edicts from around 250 BCE

[2] The notation of the International Phonetic Association (IPA) for the voiced retroflex D is ɖ, the voiceless retroflex T is ʈ, the liquid L is ɭ, and the schwa is ə.

1.1 Symbol Blocks and Intra-symbol Processing

We begin with a description of the architecture of the akshara using examples from Brahmi, the prototypical orthography of the Indic writing system. Brahmi reflects characteristics of several contemporary Indic orthographies including Kannada (Bright, 1996). Several key architectural principles can be seen in the five words written in Brahmi in Figure 1. First, each consonant-vowel unit is visually distinct from the next consonant-vowel unit (and hence the descriptor, symbol block). Second, the symbol block is not position-sensitive since its visual form is invariant wherever the symbol appears in a word. This is unlike, for example, symbol blocks in the Manchu writing system of China (see Saarela, 2014). Third, a symbol block represents either a consonant with the inherent vowel /a/ (e.g., see word final <ra> and <ha> in *hiira* and *soha*) or a consonant with a marker for vowels other than /a/ (see diacritic markers for /ii/ in *hiira*, and /i/ and /aa/ in *rimaa*). Fourth, the consonant has a perceptually prominent position in this writing system because of the use of the unmarked, and hence inherent, vowel /a/ and the way vowel markers connect to the consonant, and fifth, each symbol block comprises smaller markers (or the inherent vowel) to represent additional and more fine-grained phonological information.

Most akshara are thus analysable symbol blocks, available for intra-symbol processing. It is useful to note that these kinds of symbol deconstruction rules are mostly not relevant for some major orthographies such as the Latin-, Cyrillic- and Greek-based alphabetic systems. However, the *gulja* in Korean Hangul, the *uju* in Manchu, the Mandarin and Cantonese sets of Chinese characters, and the Japanese Kanji require similar intra-symbol decoding although the number of units that gain from intra-symbol processing differs. That is, for some symbols in each of these Asian systems, no additional linguistic information becomes available from intra-symbol analysis, but they each have a subset of symbols where intra-symbol processing aids decoding efficiency. Within the Kannada symbol set, some

symbols (e.g., consonants with the inherent vowel, the primary vowels, phonemic consonants) do not gain from intra-symbol processing but the majority of symbols do.

The mechanisms that underpin the processing of symbol blocks are under-theorised with promising lines of evidence coming from studies on writing routines when children decode (e.g., returns to insert diacritics in Tamil: Nag & Narayanan, 2019); experimental interventions (e.g., analytic approaches in Korean: Cho, McBride & Kim, 2020; copying approaches in Chinese: Wang et al., 2018); and priming studies (e.g., diacritic processing in Katakana: Kinoshita, Yu, Verdonschot & Norris, 2021; precision of representation within Korean symbol blocks: Rastle, Lally & Lee, 2019). Together, such studies may help in the understanding of the early-stage perceptual and linguistic processes involved in the use of writing systems with symbol blocks.

1.2 Orthographic Breadth and Orthographic Depth

Indic orthographies, similar to orthographies within other writing systems, also vary in orthographic breadth and orthographic depth (Share & Daniels, 2015). The Kannada orthography is more extensive in orthographic breadth than, for example, the Tamil orthography (Nag, 2007). An analysis of a collection of Kannada reading cards for primary school children found more than 700 akshara to represent Kannada phonology and several loan words. These numbers are more than found in Sinhala, the one other Indic orthography where an akshara count in children's print materials is available: language arts textbooks from grades 1 to 6 found 400 of the possible 600 or so akshara needed to capture Sinhala phonology and loan words (Wijaythilake & Parrila, 2019). This orthographic breadth indicates that users of Indic orthographies learn many symbols. Since akshara are constructed using combinatorial rules, even if mastering the large symbol set is stretched over time, the systematicity rationalises the cognitive demand (Nag, 2017). Turning to orthographic depth, Kannada is more transparent than for example, Tamil and Bengali orthographies, and this

suggests that knowing there is regularity in orthography-to-phonology mapping will make the process of learning about the Kannada orthography efficient.

Compared to orthographic depth, the study of orthographic breadth remains an understudied field although there is a growing evidence-base from Chinese and the Indic languages; for a first cross-orthographic comparison of the effects of orthographic breadth see the study of Japanese literacy acquisition by Inoue, Georgiou, Muroya, Maekawa and Parrila (2017). One implication for the teaching of extensive orthographies is the ways in which subsets of symbols are grouped and sequenced in instruction, and the finding that some aspects of the writing system are left to be inferred (e.g., Tamil: Nag & Narayanan, 2019; Bengali: Sircar & Nag, 2014).

1.3 Hidden Encoding Principles

An important design principle of the Indic writing system is the change in syllabification (re-syllabification) between the spoken word and the written word because consonants that follow a vowel must be written with a new akshara. Thus, learning to use the Indic writing system depends on gaining insights about the phenomenon of re-syllabification. In the Kannada word /pakshi/ (bird), for example, the phonological syllable units are /pak.shi/ but the akshara are <pa.kshi>. Here the post-vocalic /k/ from the first syllable /pak/ not only forms a new akshara but is also the base for including the next syllable /shi/ into the symbol block (to become /kshi/). A subtle learning demand in this context is that the printed symbol string does not reveal to the child the resyllabification process. Teaching instruction is also silent about these deep layers of representation (Nag & Perfetti, 2014). Resyllabification is thus a hidden encoding principle that must be inferred.

A final point in this brief sketch is that orthography-to-phonology mapping is mixed in Indic orthographies because the linkages occur between more than one grain size (Nag, 2017). Here, the construct of mixed mapping extends the conceptualisation of sound-symbol

mapping in the Zeigler and Goswami (2006) Grain Size Theory. In the Kannada word /pakshi/, for example, the layer of the phonological system that links to each visual feature shows that the akshara can represent more than one level of phonology. Thus in /pakshi/ the first akshara <pa> represents the body of the first syllable and the second akshara <kshi> both the coda of the first syllable and all of the second syllable. The akshara in this example therefore map onto different levels of phonology. Simpler mapping instances are also common in Kannada, such as in the /koLi/ (hen) example introduced earlier. Here, the akshara-syllable correspondence is straightforward and one-to-one (<ko.Li> - /ko.Li/). The differences in mapping seen for /koLi/ and /pakshi/ means that knowledge of variability in symbol-sound and sound-symbol correspondences is crucial for decoding efficiency in Kannada, and this holds true for most Indic orthographies.

Taken together, the learning demands to master an Indic orthography are related to visible and hidden features, illustrated here with one example each (visible: intra-symbol details, hidden: resyllabification). Given these learning demands, it is unsurprising that symbol knowledge is an important predictor of individual differences in the use of the Indic writing system (Kannada: Nag, 2007; Malayalam: Tiwari, Nair, & Krishnan, 2011; Sinhala: Wijaythilake, et al., 2019). Extending beyond the akshara writing system, symbol knowledge is also a predictor in the other Asian extensive orthographies that make a similar demand for intra-symbol processing and learning of hidden encoding principles.

2. Written Language in Children's Books

One valuable resource to understand how children come to use a writing system is to review instances where children encounter written material most often. This section describes a Kannada print corpus of children's books, before drawing out defining characteristics of the written language and thus, writing system exposure.

2.1 A Child-directed Print Corpus

The Promise Print Corpus (2020) is a collection of 15485 sequential sentences from 243 Kannada books for children between ages three and ten. These are books recommended by librarians, teachers, and parents. Here, the term ‘book’, is used to refer to books with a single story, anthologies, drama scripts, textbooks and reading cards with brief narratives. The predominant genre in the corpus is contemporary fiction with a small selection of mythical stories, moralistic folktales, fables, and condensed episodes from Indian epics. A very small set is non-fiction and on science themes. Other print materials children encounter such as children’s magazines, comics, flashcards, information posters and instruction sheets in school are not in the print corpus. Nonetheless, the coverage statistics (reported next) suggests the Promise Print Corpus (2020) is a valuable source to understand the nature of encounters a novice learner is likely to have with written language.

The corpus contains the works of 126 writers, three of whom are award-winning authors. As is common within the publishing industry in south Asia, several books are translations, usually from another Indian language. The translations come from 49 translators with five translating between six and ten books, and two translating 15 and 22 each. Several books do not name the illustrator but where available, the stories are drawn by 75 illustrators. This number of contributors helps ensure the written language in the corpus is not skewed to a particular type of language usage. Coverage by year of publication is also important since the quantity and content in child-directed print has changed in the last decades both in terms of investment by publishing houses in children’s books and in topic range and narrative styles (e.g., Prabhu, 2019; Shankar, 2018). The corpus carries 13% of books published before 2000, 45% between 2001-2010, and 40% since 2011. Two percent of publications in the corpus did not print the publication year but book owners recalled the book was acquired before year 2000. The corpus has books from fifteen publishing houses, with eleven percent of books listed as bestselling by the publisher.

Books targeting the three-to-eight-year-old child are over-sampled in the corpus, making it especially suitable to understand task demands on emergent and beginning readers. There are 88 read-aloud and picture books for three-to-five-year-olds, 118 books for early readers in the six-to-eight age band and 36 books for the nine and ten age band. Discounting reading cards and anthologies in the set, books for older children show the expected increase in page-, sentence- and word-count. The number of pages in books for the youngest age band range from eight to 27 with many pages being wordless as expected in picture books and read-aloud books. In books for beginning readers, the number of pages range from seven to 39, and eleven to 51 in books for nine- and ten-year olds. The average number of sentences increase across age bands ($M_{3\text{-}5 \text{ age band}}$ = 21.64 (SD = 15.00), $M_{6\text{-}8 \text{ age band}}$ = 38.88 (SD = 36.49), and $M_{9\text{-}10 \text{ age band}}$ = 157.83 (SD = 216.67)) and similarly the number of words ($M_{3\text{-}5 \text{ age band}}$ = 119.91 (SD = 87.16), $M_{6\text{-}8 \text{ age band}}$ = 240.13 (SD = 229.70), and $M_{9\text{-}10 \text{ age band}}$ = 1142 (SD = 1353.27)). Although neither number of pages nor the sentence- and word-count in a book can be taken to judge quality of language, these structural indicators provide an estimation of quantity of exposure to written language. The descriptive data also shows variation within each age band for number of sentences and words per books. This again implies that even if age banding captures a range in structural characteristics of books it is still a useful proxy for the nature of exposure that children have to the writing system.

2.2 The Written Language: Key Characteristics

A descriptive analysis of the written language in the Promise Print Corpus (2020) was initiated through a lexical search for one key word. This section reports the findings. For the lexical search, sentences that included the early acquired word /aane/ (‘elephant’) were identified. This bisyllabic word was chosen because it appears often in picture books, animal charts and phonic drills for pre-schoolers (Indic literacy instruction begins with the vowels /a/ and /aa/; see Nag & Perfetti, 2014). The akshara in /aane/ are, however, not amongst the most

frequent in the corpus. While the akshara frequency of <aa.ne> is 3688 and 2303 respectively, the akshara frequency for another co-occurring bisyllabic word <va.na> (forest) is 5815 and 7758 respectively. The word /aane/ is also interesting because its inflected and derived forms reflect well the morphology of Kannada. Five sentences from the pool of /aane/ sentences were picked next. Each sentence was from a different book, and each book from a different publisher and with a different year of publication (1972, 1989, 1993, 2006, 2016). Two sentences each are from books for three-to-five-year-olds and nine- and ten-year olds, and one from six-to-eight-year-olds. The sentences are as follows (the lexical search word and relevant gloss are highlighted in bold):

*S1. ondu kaaDinalli **aane** mattu niir**aane** oTTige vaasisuttiddavu.*

ondu kaaD(i)-(n)alli **aane** mattu niir**aane**

a forest-LOC **elephant** and **hippopotamus** (water elephant)

oTTige vaas-isu-tt-idd-avu

together live-VRB-PRES-IR.PST-3N.PL

(In a forest, an elephant and a hippopotamus lived together.)

*S2. **aane**, cirate, karaDi haaguu jinke ellariguu cinte aayitu.*

aane, cirate, karaDi haaguu jinke ell-ar-ig-uu

Elephant leopard bear and deer all-PL-DAT-CLT.INCL

cinte aayitu

worry happen-IR.PST

(The elephant, leopard, bear and deer were all worried.)

*S3. aa **aaneya** kooredaaDegaLu muuru naalku aDi uddaviddavu.*

aa **aane-(y)a** kooredaaDe-gaLu muuru naalku aDi udda-(v)idd-avu

That **elephant-GEN** tusk-PL three four foot long-IR.PST-3N.PL

(That elephant's tusks were three to four feet long.)

S4. ***aanegaLannu*** *paLagisuvavanobba ii* ***aaneyannu*** *nooDikoLLuttidda.*

Aane-gaL-annu paLagisu-(v)avan-obba

elephant-PL-ACC cause to become trained-he-3M.SG-a (a mahout)

ii **aane-(y)annu** nooD-i-koLLu-tt-idd-a

this **elephant-ACC** Look-PST.PPL-BEN-PRES-IR.PST-3M.SG

(A mahout looked after this elephant.)

S5. modalu … aane hola, kumbaaragunDi haaguu bannuurige hoodaru.

modalu aane hola kumbaaragunDi haaguu bannuur-ige

First… **Aane Hola (proper noun)** Kumbaaragundi and Bannuur-DAT

hood-aru

go-IR.PST-3H.PL

(First we went to… Aane Hola, Kumbaragundi and Bannur …)

Several key aspects of the nature of written language in children's books are captured in these sentences and each signal the morphological, orthographic and phonological skills needed to engage with the writing system and read with comprehension. Five aspects among these are discussed next.

2.2.1 The Language is Morphologically Rich. The Kannada sentences show a range of inflectional morphology and derivational morphology. There are several bound morphemes, and the case relations in content words are grammaticalized with postpositions. These are seen in the case inflection and the markers to distinguish number and gender in the corpus search word, /aane/ (S1-S4). Another noun example is the commonly co-occurring

word /kaaDu/ (‘forest’): *kaaD.inalli*, ‘in a forest’ (LOC: locative case); *kaaD.ige*, ‘to the forest’ (DAT: dative case), and *kaaD.ina*, ‘of the forest’ (GEN: genitive case). While nouns are multimorphemic, the more densely inflected word in most sentences is the verb (see panel two in Table 1).

Compound constructions are also common in Kannada. The word /aane/, for instance, appears as the compound word *niir.aane* (‘water elephant’, ‘hippopotamus’, S1). Elsewhere in the corpus, further instances of /aane/ compounds were *kaaD.aane* (‘wild elephant’) and *heNN.aane.gaLu* (‘female elephants’). Another example of noun + noun compound is *paLagisu-(v)avan-obba* (‘one who looks after’, ‘the mahout’, S4). Like noun+noun compounds, serial verbs are also present in children’s books (see panel two in Table 1). From the perspective of what children must bring to the task of reading, these words define the linguistic knowledge needed for comprehension of the multi-faceted nature of morphological information. Literacy instruction in Kannada rarely, however, focuses on ensuring strong oral language foundations. The assumption instead is that language learning, especially among native speakers, will occur spontaneously and to similar levels of advanced linguistic knowledge for all children. Systematic analysis of children’s oral language however does not confirm this assumption (e.g., Nag, Snowling & Mirkovic, 2018a), signalling the need for interventions that develop children’s broader oral language skills. One component of a Kannada oral language intervention may be analytic activities about the structure of multimorphemic words. Such an approach is in line with the growing consensus that morphological awareness plays a strong role in learning to use a writing system irrespective of writing system characteristics (e.g., opaque, alphabetic English and transparent, alphabetic Greek: Manalotis, Georgiou, Inoue, & Parrila, 2019).

Another implication of the repeated appearance of multimorphemic words is the orthographic knowledge needed to decode. The next sections examine inferences from the five sentences about the orthographic task demand present in children's books.

2.2.2 Morpho-phonological Changes Occur in Many Words. Hidden within word formation processes in Kannada are changes in phonological values that occur at the boundary between morphemes. The phonological modifications to combine morphemes are governed by a set of *sandhi* rules. An example of a morpho-phonological change is the dropping of the word-final /e/ from the root word /beLigge/ ('morning') when a genitive case marker and the number and gender markers /-ina/ are added, *beLiggina*. Similar omission of the word-final vowel is seen for /kaaDu/ in the previous section. Also common are phonological additions, with what is affixed depending on boundary characteristics. For example, for the suffix /-inda/ (the instrumental-case + number + gender marker), the boundary change adds /-y-/ for one word and /-in-/ for another: *shaaley<u>y</u>inda* ('from school', shaale-(y)inda) and *dikk<u>in</u>inda* ('from the direction', dikku(i)-(n)inda), and the word-ending vowel is dropped in the second but not the first example.

Omission and addition of phonological units are examples of word level morpho-phonology processes and they appear in all five sample sentences under study. There are several other morphophonological processes within the Kannada language (see Krishnamurti, 2003; Ranganatha, 1982; Sridhar, 1990). For a beginning reader, these morpho-phonological changes demand knowledge of word formation rules, orthographic knowledge and phonological processing skills.

2.2.3 The Word Lengths are not that Short. The word length of content words is relatively long in Kannada and one reason for this is the agglutinative nature of the language. The number of akshara per word for a sub-set of words from sentences 1 to 5 is given in Table 1. Within this set, verbs have five and six akshara while nouns range from two to eight

akshara. There are a few one akshara words but this is a rarity. Thus, the sampled sentences show that recognition and blending of three to six akshara words is quite routine in children's books. Since akshara are symbol blocks, a three to six akshara word would, at a minimum, translate into a six to twelve letter word if rendered in an alphabetic orthography.

Comparing word lengths seen in agglutinative Kannada to other languages can help further understand the skills children bring to the task of using this writing system. In alphabetic English for example, the classic study on length statistics by Miller, Newman & Friedman (1958) is a good starting point. Using chapters from a religious book, a scholarly book and a news journal (i.e., an adult-directed print corpus), the average word length for content words in English was 6.47 letters. Miller and colleagues pointed to how word lengths in the morphologically sparse English were 'more nearly random' in comparison to synthetic languages such as Russian, where they proposed word lengths to be 'constrained by morphological rules' (pg. 374). This early statement about word lengths in agglutinative languages remains valid with several lines of research examining how readers disambiguate the affixes and stems in varying word lengths (e.g., eye tracking studies with Finnish 6 to 9 letter and Uighur two to 21 letter multimorphemic words: Hyönä, Yan & Vaino, 2018; Yan et al., 2014 respectively). Similar lines of research are needed for Kannada, and the other Indic orthographies, to clarify the underpinnings of akshara-by-akshara decoding as well as the higher order processing of the root-and-affix components in words.

2.2.4 The Symbol Set includes Low and High Frequency Akshara. A symbol-level review of the words in sentences 1 to 5 show that all akshara types are present in children's books. This includes words with primary vowels (e.g., *adi, aane, ii*), words with the inherent vowel (underlined) and other vowels (*hola, cirate, karadi*), words with the coda marker for /n/ (*jinke, cinte, ondu*) and words with the complex akshara that carry a consonant cluster with either an inherent vowel or another vowel (*mattu, naalku, uddaviddavu*).

Importantly, some akshara are frequent and others lower in frequency such as in the number words for 'three' (/muuru/) and 'four' (/nalku/) in S3 above: the frequency count of the akshara in <muu.ru> is 384 and 5004 and of those in <na.lku> is 2368 and 48 respectively. For a child learning to use the writing system, essential orthographic knowledge includes knowing different akshara types and both higher and lower frequency akshara.

2.2.5 The Mapping of Sound-to-symbol is mixed. Finally, many words in the five example sentences have a one-to-one correspondence between the akshara and the phonological syllable. Examples include <aa.ne>, <mo.da.lu> and <koo.re.daa.De.ga.Lu>. In parallel, the presence of resyllabification in Indic orthographies means words also have mixed mappings. The variety in mapping is shown in Table 1 and the akshara with cross-syllable mapping are underlined (see column two). Importantly, an examination of the instances of akshara to coda + next syllable mapping shows that these occur substantially more often in word internal positions and are most often at morpheme boundaries. This has implications for children's use of decoding skills as well as exposure to the variety of orthographic representations in the writing system.

Table 1

Words with word length and akshara-syllable mapping

Word	Word length (number of akshara* with mismatched mapping underlined)	Meaning
Nouns with root word /aane/ ('elephant')		
aaneya	aa.ne.ya (3)	elephant's
aaneyannu	aa.ne.ya.nnu (4)	an elephant
aanegaLannu	aa.ne.ga.La.nnu (5)	elephants
niiraane	nii.raa.ne (3)	hippopotamus (water elephant)
aane hola	aa.na ho.la (2, 2)	Elephant Field (a village name)

Verbs		
vaasisuttiddavu	vaa.si.su.tti.dda.vu (6)	lived
uddaviddavu	u.dda.vi.dda.vu (5)	was long
nooDikoLLuttidda	noo.Di.ko.LLu.tti.dda (6)	looked after
Other words		
aa	aa (1)	that
Ii	ii (1)	this
aDi	a.Di (2)	feet (metric)
muuru	muu.ru (3)	three
naalku	naa.lku (2)	four
kooredaaḍegaLu	koo.re.daa.De.ga.Lu (6)	tusks
paLagisuvavanobba	pa.La.gi.su.va.va.no.bba (8)	mahout

Note: * a mapping is considered mismatched when there is no one-to-one correspondence between akshara and the phonological syllable.

3. Developmental Evidence on Learning to Read

In this section, findings about written language in children's books will be examined against the available evidence on learning to read in the Indic orthography of Kannada. Reviews of this evidence-base are already available (Nag, 2017a, 2017b) and will not be repeated here. Instead, the key trends from the exploratory analysis of a child-directed print corpus will be matched to available evidence to identify divergence in what is examined in current research and what is expected within children's books. To recap, the written language in children's books reflects the rich morphology of Kannada and in the process introduces the extensive symbol set within word lengths often three to six akshara long, with mapping of the akshara to a syllable, either the body or the coda of a syllable, or across two syllables, and in words that may also include morpho-phonological changes.

3.1 **Symbol Knowledge**

Children with a greater knowledge of the writing system know which intra-symbol cues to attend to and what they signal about symbol identity (e.g., Nag, 2014; Nag &

Narayanan, 2019). In addition, those who know more symbols can decode more words, with the available evidence focused on the effects of akshara type (Nag, 2007; Tiwari et al, 2011; Wijayathilake & Parrila, 2019) and type of akshara-syllable mapping (Nag, 2014; Sircar & Nag, 2014). Children who recognise that orthographic representation is for more than one phonological grain size, move quickly to understanding the writing system. The mechanisms underpinning akshara learning mirror the frequency-and-regularity effects on word learning (e.g., see Harm & Seidenberg, 2004) but, first reported in Nag (2014), instances of encounter may be low for any single akshara while encounters with a design principle of the writing system is high. This area of low frequency but high regularity effects on orthographic learning is understudied, and a child-directed print corpus may be a valuable resource to begin examining this field systematically.

Other symbol-level characteristics that have received research attention include frequency, consistency, phonological confusability, complexity of visual features and the visuo-motor demands of writing individual symbols. A methodologically robust way to assess the inputs of each of these symbol characteristics is to assess if they remain a predictor even after the other features have been included in a statistical model. For the Kannada orthography, all these features were important in the earliest stages of symbol learning, with the visual complexity of the symbol remaining a predictor of symbol recognition even after frequency and phonological confusability were considered (Nag, et al. 2014). In parallel, child-level skills that better support symbol learning included visual processing skills to analyse the visual features of the symbols, phonology-linked skills, and vocabulary. The finding from this exploratory child-directed print analysis, however, suggests that there is a complex series of demands from not only phonology and orthography but also morphology to read words in children's books. A gap in the current evidence-base for Kannada, but also other Indic orthographies, is to explain the role of morphology and morpho-phonological

processes in symbol processing (e.g. to disambiguate visual word recognition if there is a low frequency symbol that is not recognised).

3.2 Visual Word Recognition Skills

Decoding a word symbol-by-symbol using symbol knowledge is a clear case of using knowledge of the writing system to support reading. But broader language skills are also important. The role of morphology is important for a variety of reasons including for informed guessing of the meaning of unfamiliar words, supporting new word learning by providing cues, and providing morpheme-level clues on how to read an unfamiliar word. The data analysed with sentences from the child-directed print corpus shows this knowledge is essential to successfully decode the kinds of words found in 'real world' encounters. It is known that morphological knowledge has a somewhat different role in different languages depending on the characteristics of the orthography in use. In opaque orthographies, morphological knowledge facilitates decoding of most words while in transparent orthographies the decoding facilitation is reserved for low-frequency words. In the Indic orthography, and based on the words found in the example sentences in children's books, morphological knowledge for visual word recognition may be needed for disambiguating components in a multimorphemic words at two levels: when decoding processes encounter low frequency akshara or there is a mismatch in the default akshara-syllable mapping rule. Accuracy at the morpho-phonological boundary has been shown to be predicted by phonological awareness skills in a language production task (Nag et al., 2018a), and inflection knowledge in a reading comprehension task (Nag & Snowling, 2011). More such research is needed because the demand in real-world reading is clearly for competencies with multimorphemic words and morpho-phonological awareness.

In conclusion, models of orthographic learning in an Indic writing system have tended to focus on intra-symbol cues, the size of the symbol set, and the nature of sound-symbol

mapping (e.g., Nag, 2007, 2014), although empirical studies on each remain scarce. Aside from these skills, the task demands of using the writing system as implicit in children's books show the substantial demand for visual word recognition of multimorphemic words. Since children encounter an ever-expanding variety of such words in the books they read, it is essential to mount systematic studies on the development of this area of skills in Kannada, and the Indic languages more generally. At the pedagogical level, research using child-directed print corpora can additionally suggest what, when and how to support children's literacy learning; at a methodological level, this exploratory study shows the limitations of building models of literacy development when real-world encounters with a writing system are not adequately taken into account.

References

Bright, W. (Ed.). (1992). *International Encyclopaedia of Linguistics*. New York, NY: Oxford University Press.

Bright, W. (1996). Kannada and Telugu writing. In P.T. Daniels & W. Bright, *The World's Writing Systems*, 413-419. Oxford, UK: Oxford University Press.

Cho, J.R., McBride, C. & Kim, B. (2020). Effective teaching instructions for Hangul learning among Korean kindergartners. *Reading and Writing* 33: 1791–1808.

Coates, J. (2017). *The Effect of Story Grammar Contained in Wordless Picture Books: A Study of Children's Narrative*. Unpublished Masters Thesis. University of Oxford, UK.

Daniels, P.T. (1990). Fundamentals of grammatology. *Journal of the American Oriental Society* 110: 727-731.

Harm, M., & Seidenberg, M.S. (2004). Computing the meanings of words in reading: Division of labor between visual and phonological processes. *Psychological Review* 111: 662-720.

Hayes, D.P., & Ahrens, M.G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of Child Language* 15: 395–410.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behaviour and Brain Science* 33(2-3): 61-83.

Hyönä, J., Yan, M., & Vainio, S. (2018). Morphological structure influences the initial landing position in words during reading Finnish. *Quarterly Journal of Experimental Psychology* 71(1): 122–130.

Inoue, T., Georgiou, G.K., Muroya, N., Maekawa, H. and Parrila, R. (2017). Cognitive predictors of literacy acquisition in syllabic Hiragana and morphographic Kanji. *Reading and Writing* 30: 1335–1360.

Kinoshita, S., Yu, L., Verdonschot, R.G., & Norris, D. (2021). Letter identity and visual similarity in the processing of diacritic letters. *Memory and Cognition* 118(3): 1–11.

Krishnamurti, B. (2003). *The Dravidian Languages*. Cambridge: Cambridge University Press.

Manolitsis, G., Georgiou, G.K., Inoue, T. & Parrila, R., (2019). Are morphological awareness and literacy skills reciprocally related? Evidence from a cross-linguistic study. *Journal of Educational Psychology* 111(8): 1362–1381.

Massaro, D.W. (2015). Two different communication genres and implications for vocabulary development and learning to read. *Journal of Literacy Research* 47: 505–527.

Miller, G.A., Newman, E.B. & Friedman, E.A. (1958). Length-frequency statistics in written English. *Information and Control* 1: 370–389.

Montag, J.L. (2019). Differences in sentence complexity in the text of children's picture books and child-directed speech. *First Language* 39(5): 527-546.

Montag, J.L., Jones, M.N., & Smith, L.B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science* 26(9): 1489-1496.

Nag, S. (2007). Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading* 30: 7-22.

Nag, S. (2014). Akshara-phonology mappings: The common yet uncommon case of the consonant cluster. *Writing Systems Research* 6: 105–119.

Nag, S. (2017a). Learning to read alphasyllabaries. In K. Cain, D. Compton, & R. Parrila (eds.), *Theories of Reading Development*. Netherlands: John Benjamins.

Nag, S. (2017b). Learning to Read Kannada and Other Languages of South Asia. In L. Verhoeven & C. Perfetti (eds.), *Learning to Read across Languages and Writing Systems*, 82-103. Cambridge: Cambridge University Press.

Nag, S. & Narayanan, B. (2019) Orthographic knowledge, spelling and reading development in Tamil: the first three years. In M. Joshi & C. McBride (eds), *Handbook of Literacy in Akshara Orthography*. Dordrecht: Springer.

Nag, S. & Perfetti, C.A. (2014). Reading and writing: Insights from the alphasyllabaries of South and Southeast Asia. *Writing Systems Research* 6(1): 1-9.

Nag, S. & Snowling, M.J. (2011). Reading comprehension, decoding and oral language, *The EFLU Journal*, *English and Foreign Languages University* 2(2): 75-93.

Nag, S., Snowling, M.J. & Mirkovic, J. (2018a). The role of language production mechanisms in children's sentence repetition: Evidence from an inflectionally rich language. *Applied Psycholinguistics* 39(2): 303-325.

Nag, S., Snowling, M.J., Quinlan, P. & Hulme, C. (2014). Child and symbol factors in learning to read a visually complex writing system. *Scientific Studies of Reading* 18: 1-16.

Nag, S., Vagh, S.B., Dulay, K.M. & Snowling, M.J. (2018b). Home language, school language and children's literacy attainments: a systematic review of evidence from low- and middle-income countries, *Review of Education* 7: 91-150.

Nag, S., Vagh, S.B., Dulay, K.M. & Snowling, M.J., Donolato, E. & Melby-Lervag, M. (in preparation). Home attributes that relate to language and literacy attainments: A systematic review of studies from low- and middle-income countries.

Pearce, W.M. (2003). Does the choice of stimulus affect the complexity of children's oral narratives? *Speech Language Pathology* 5(2): 95-103.

Prabhu, R.D. (2019.7.14). Children's literature in India undergoes revolution as publishers experiment with regional languages, genres. *First Post.* *https://www.firstpost.com/living/childrens-literature-in-india-undergoes-revolution-as-publishers-experiment-with-regional-languages-genres-6943791.html*

Promise Print Corpus (2020). *A Child-Directed Print Corpus of Kannada Books.* Bangalore, India: The Promise Foundation.

Ranganatha, M.R. (1982). *Morphophonemic analysis of the Kannada language: Relative frequency of phonemes and morphemes in Kannada*. Mysore: Central Institute of Indian Languages.

Rastle, K., Lally, C., & Lee, C.H. (2019). No flexibility in letter position coding in Korean. *Journal of Experimental Psychology: Human Perception and Performance* 45(4): 458–473.

Saarela, M. S. (2014). The Manchu script and information management: Some aspects of Qing China's great encounter with alphabetic literacy. In B.A. Elman (Ed.), *Rethinking East Asian languages, vernaculars, and literacies, 1000-1919*, 169–197. Leiden: Brill.

Sircar, S. & Nag, S. (2014). Akshara-syllable mappings in Bengali: a language-specific skill for reading. In H. Winskel & P. Padakannaya (Eds), *South and Southeast Asian Psycholinguistics,* 409-425. Cambridge: Cambridge University Press.

Shankar, K. (2018.2.17). Bright prospects for children's publishing in India. In *The Hindu*, https://www.thehindu.com/books/the-prospects-of-childrens-publishing-in-india-are-brightening/article22771146.ece

Share, D.L. & Daniels, P.T. (2015). Aksharas, alphasyllabaries,abugidas, alphabets and orthographic depth: Reflections on Rimzhim, Katz and Fowler (2014), *Writing Systems Research* 8(1): 17-31.

Shapiro, L.R., & Hudson, J.A. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives. *Developmental Psychology* 27(6): 960-974.

Sridhar, S.N. (1990). *Kannada.* London: Routledge.

Wang, Y., McBride, C., Zhou, Y., Joshi, M., & Farver, J. (2018). Becoming literate in Chinese: A comparison of native and non-native speaking children. *Journal of Research in Reading* 41: 511–524.

Wijaythilake, M.A.D.K., Parrila, R., Inoue, T. & Nag, S. (2019). Cognitive predictors of word reading in Sinhala. *Reading and Writ*ing 32: 1881–1907.

Wijaythilake, M.A.D.K. & Parrila, R. (2019). Reading and writing Sinhala. In R.M. Joshi & C. McBride (eds), *Handbook of Literacy in Akshara Orthography*, 195-216. Dordrecht: Springer.

Yan, M., Zhou, W., Shu, H., Yusupu, R., Miao, D., Krügel, A., & Kliegl, R. (2014). Eye movements guided by morphological structure: Evidence from the Uighur language. *Cognition* 132: 181–215.

Ziegler, J. C., & Goswami, U. (2006). Becoming literate in different languages: similar problems, different solutions. *Developmental Science* 9: 429-436.