Looking at children's books to fill gaps in the science of reading

Sonali Nag

University of Oxford

Psychonomic Society Symposium: Cognition in the "real-world": How, Why and What next?

Experimental Psychology Society (EPS), London Meeting, 6th January 2023

Talk Together

Supporting Oral Language Development

Department of Education, University of Oxford, Experimental Psychology, Medical Sciences Division

The Promise Foundation — For Mental Health, Education and Potential Realisation

Georgia State University

UKRI Economic and Social Research Council

GCRF

Web: http://talktogether.web.ox.ac.uk

Twitter: @talktogproject

In this presentation

- Look through the lens of writing systems variation
- Highlight the dramatic skew in the current evidence-base
- Consider potential research infrastructure to support work in understudied languages

The science of reading

Deeply rooted in a cognitive-linguistic view of literacy learning

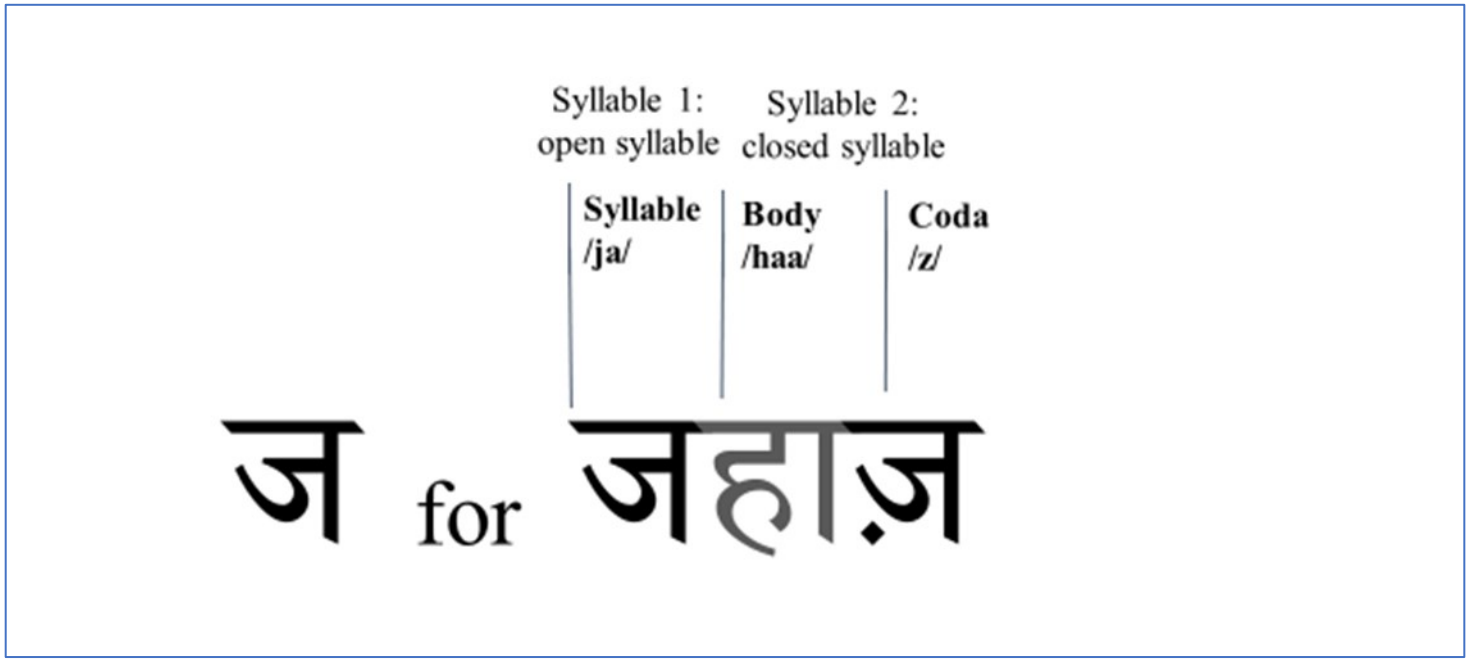
- The oral language foundations for reading development
- The mapping principles of written language systems

- Developmental trajectories
  - Effortful computation → automatized retrieval
  - Lower-order rules → higher-order principles

- Cognitive-linguistic outcomes affected by their ecological embedding
  - Literacy practices
  - Language of instruction
  - Home language and literacy environments

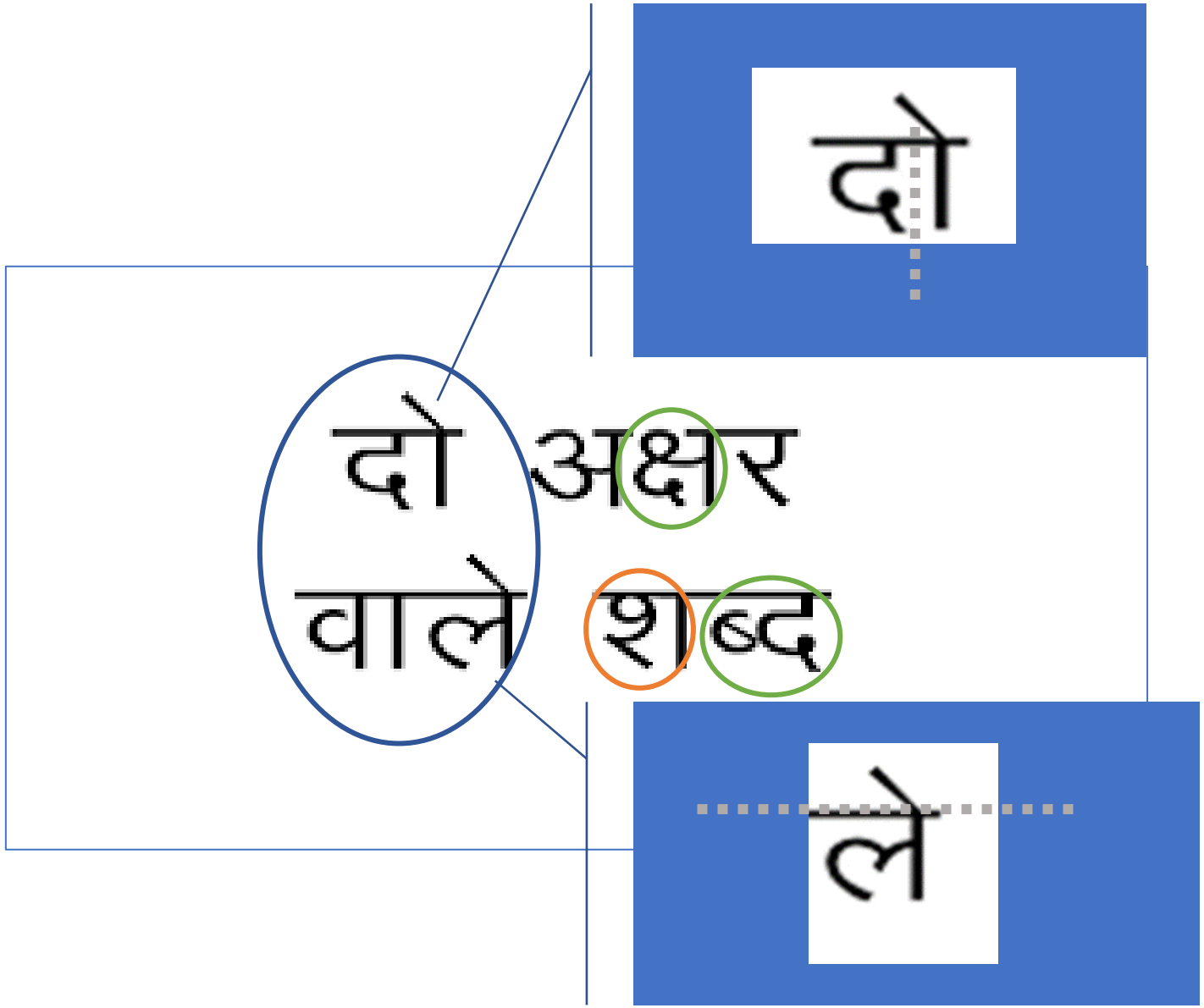
Writing systems variation

- **The mapping principle: Indic akshara system (Hindi)**



Nag, S. (2022). Reading the Akshara System**.** In M. Snowling, C. Hulme & K. Nation (Eds.) *The Science of Reading: A Handbook*. John Wiley & Sons..

- Ca
- CV
- CCV



Writing systems variation

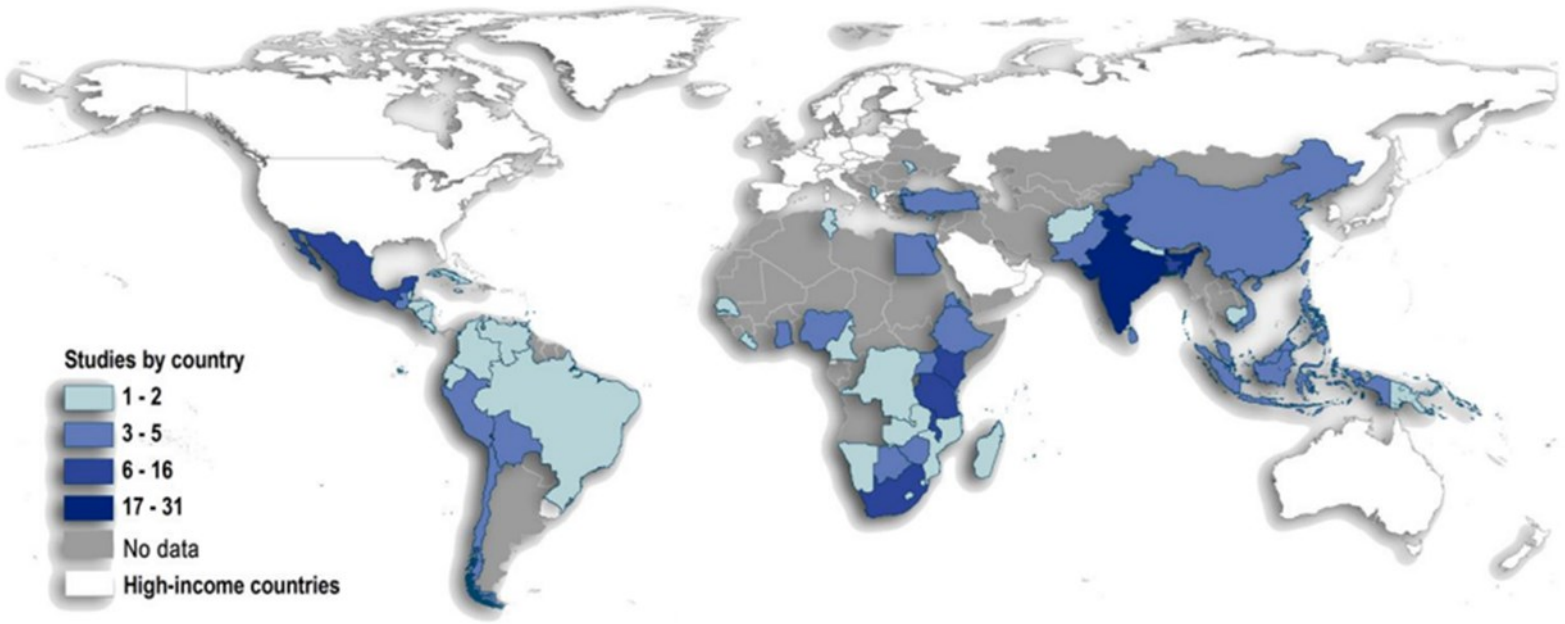
Example parameters

- Inventory size
  - Frequency
- Transparency
  - Letter-sound, sound-letter
- Visual complexity
  - Intra-symbol processing

Theoretical implications

- Role of semantic bootstrapping
- In symbol-level decoding
- Visual confusability

Current evidence base

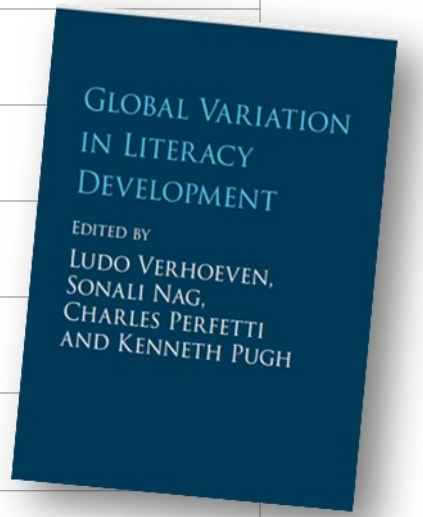


- Arabic*
- Bahasa Indonesian#
- Bahasa Melayu #
- **Bengali, Kannada & Odia***
- English & Spanish
- Herero & Zulu #
- Kiswahili*
- **Tigre & Tigrinya***
- Turkish #

Nag, Chiat, Torgerson & Snowling (2014)
Literacy, Foundation Learning and Assessment in Developing Countries. DFID.

Current evidence base

Global Variation in Literacy Development
• Sociocultural variation in literacy Northern America
• Literacy variation in South America
• Postcolonial literacy in the Caribbean
• Literacy development in Europe
• Challenges for literacy in Russia
• Literacy acquisition and language of instruction in Sub-Saharan Africa
• Literacy and diversity in multilingual India
• Reading acquisition in South-East Asia
• Literacy and linguistic diversity in Australia



Verhoeven, Nag, Perfetti & Pugh (Eds) (in press).
Global Variation in Literacy Development. Cambridge University Press.

Reflections in the language acquisition literature

Paradis, J. (2022). What can journals do to increase the publication of research on the acquisition of understudied languages? A commentary on Kidd and Garcia (2022). *First Language*, *42*(6) 794–798.

Current evidence-base

- Focus on research infrastructure to support work in understudied languages

- Examples of research infrastructure
  - The sketch acquisition project (Hellwig et al., 2021)
  - Tool for intergenerational transmission assessment protocol (Deen et al, 2016)

Two practical propositions

1. Child-directed print corpora in understudied languages
2. The mapping of the psycholinguistic properties of such corpora

The opportunity: more book titles!

- in languages traditionally underrepresented in the publishing industry.

The challenge: can less give more?

- approaches to corpora construction understudied
- mainly dependent on manual tagging

The practical usefulness of a small-sized corpus

What, if any, is the equivalence in word-level characteristics between a larger corpus and a smaller corpus?

- A cross-corpora analysis to compare word length
- A developmental analysis of orthographic representation and orthographic diversity by book level

Case study: Kannada

- **Corpus description**
  - Books for 3-10 year olds
  - Picture books, story collections, folktales, chapter books, non-fiction books, learner's materials and translated works
  - Book lengths range from 7 to 1754 sentences

  - **Bigger corpus:** 151,249 words from 24,375 sentences from 402 books

  - **Smaller corpus:**
  - Non-sequentially sampled corpus: 17,584 words from every 10^{th} sentence from the 402 books

John, S., Agarwal, A. & Nag, S. (in preparation). Examining a short-version of a child-directed print corpus in an understudied language: A cross-corpora and psycholinguistic analysis

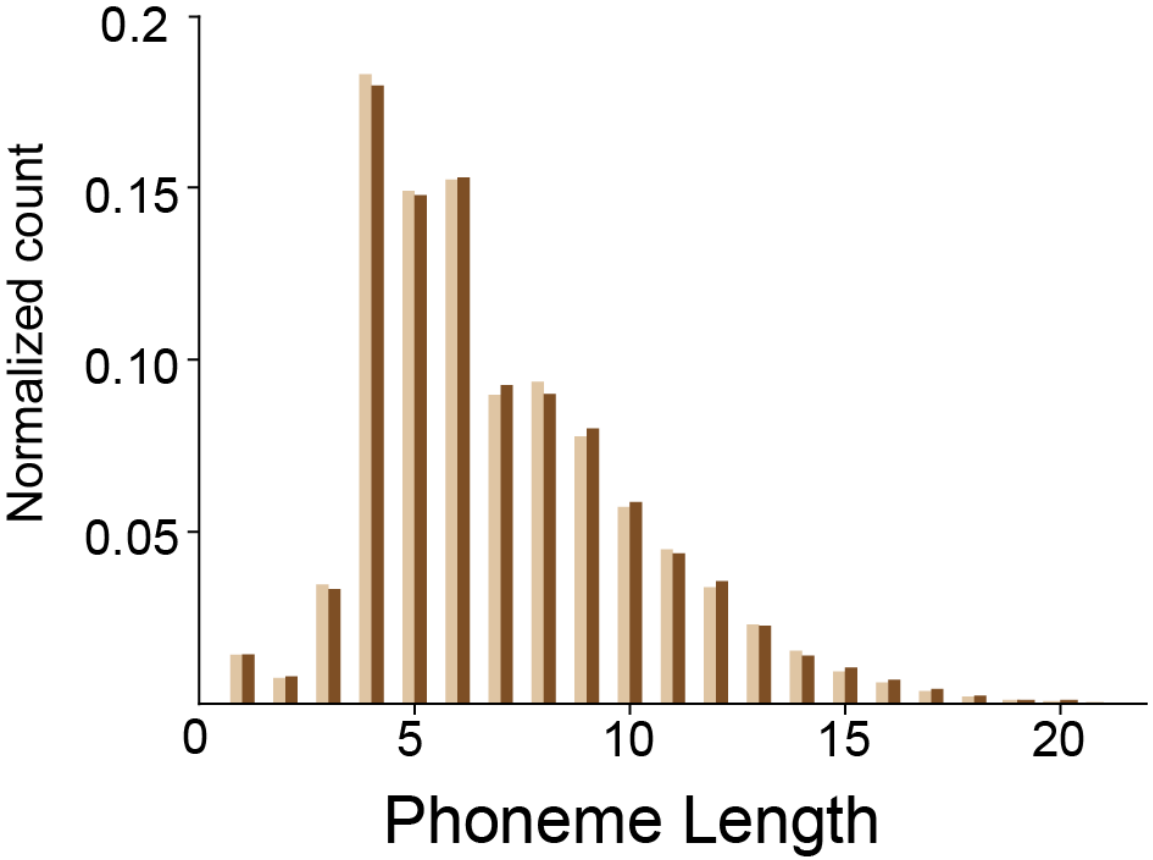
Selective non-sequential sampling protocol

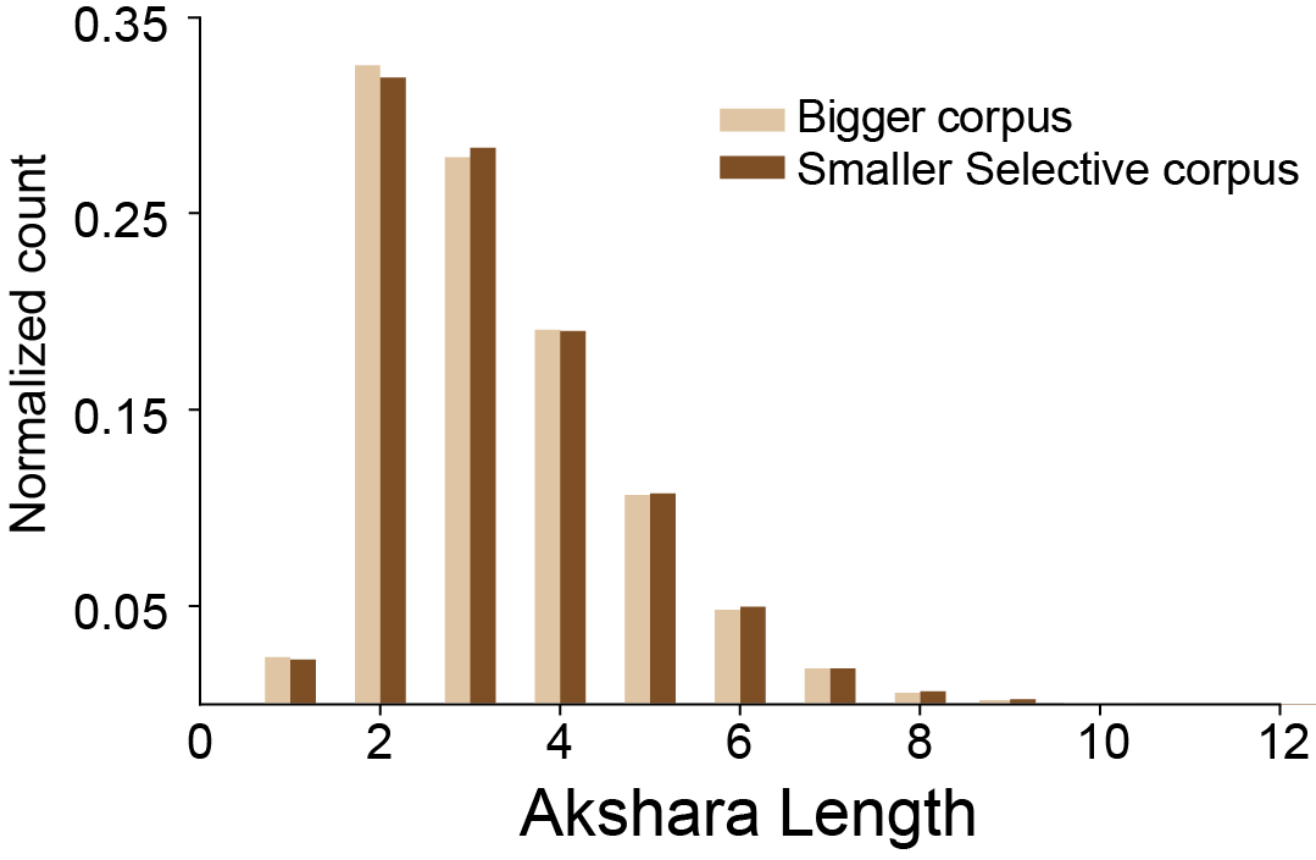
- Books with less than 10 sentences
  - 1 sentence randomly chosen
- Books with more than 10 sentences but less than 32 pages
  - every 10^{th} sentence
- Books with more than 32 pages
  - a text track of 1500 words was randomly selected and then every 10^{th} sentence

- To improve random selection
  - sentence selection began from one of the first three sentences identified through a random number generator.
  - sentence 1, 11, 21…, or 2, 12, 22…, or 3, 13, 23...

Nag, S., Dulay, K., Arulmani, G & Ocampo, D. (2021). TalkTogether Child Directed Print Corpus. Technical Report for Kannada and Filipino.

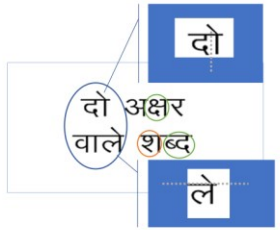
Word length





p = 0.07, unpaired t-test

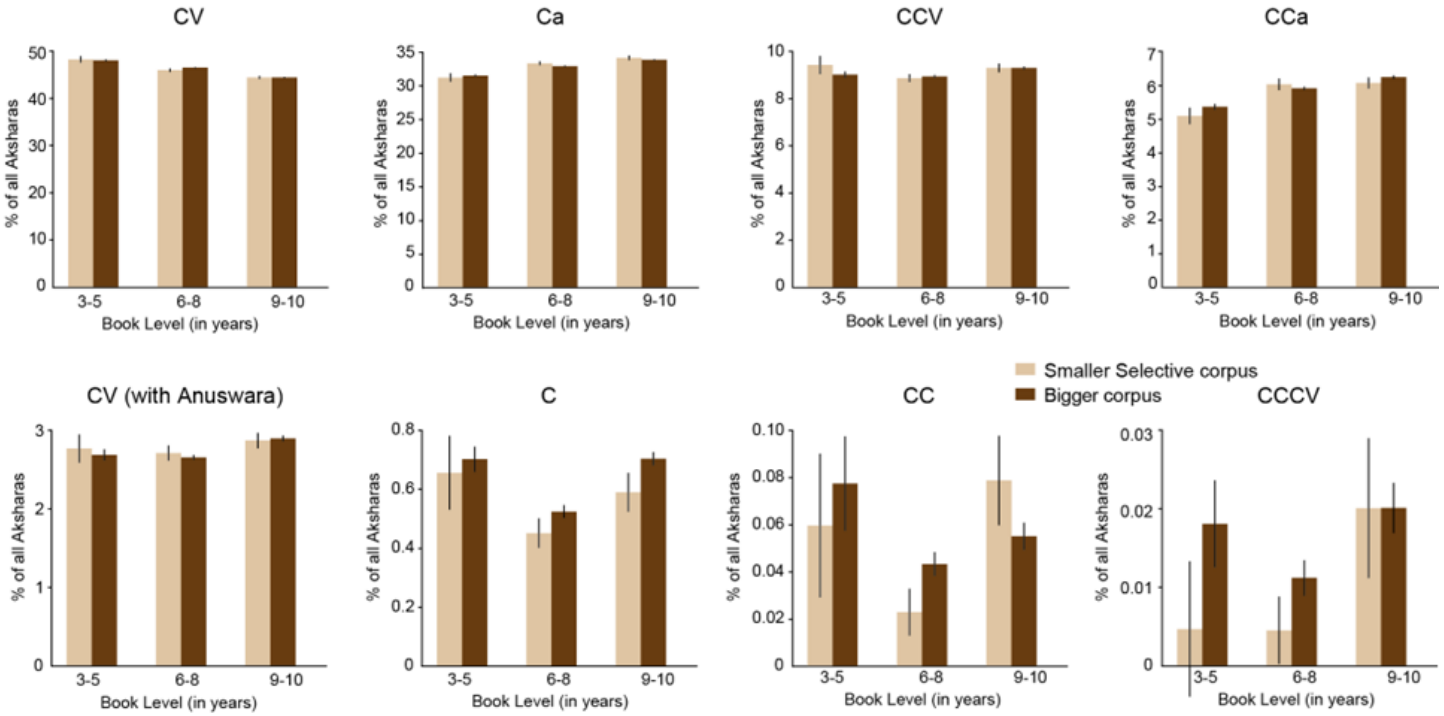
Orthographic representation



Occurrence statistics of
8 akshara types per corpus

By book level

- Since word counts more in older books raw count not appropriate; normalised frequency count used
- 100 iterations with randomly selected words and repeated replacement before next random selection



Orthographic diversity

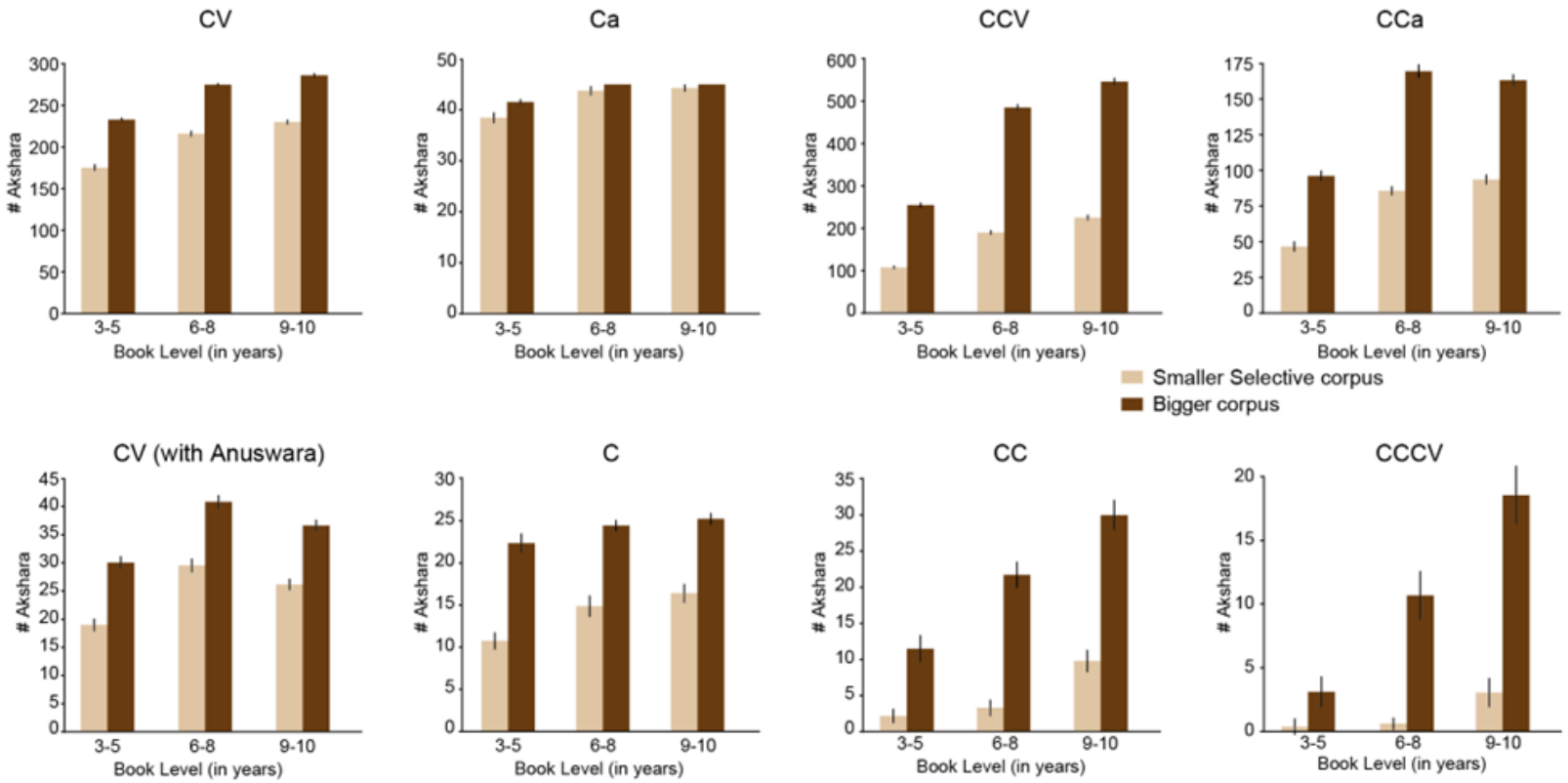
Akshara diversity:

number of unique akshara in each book level

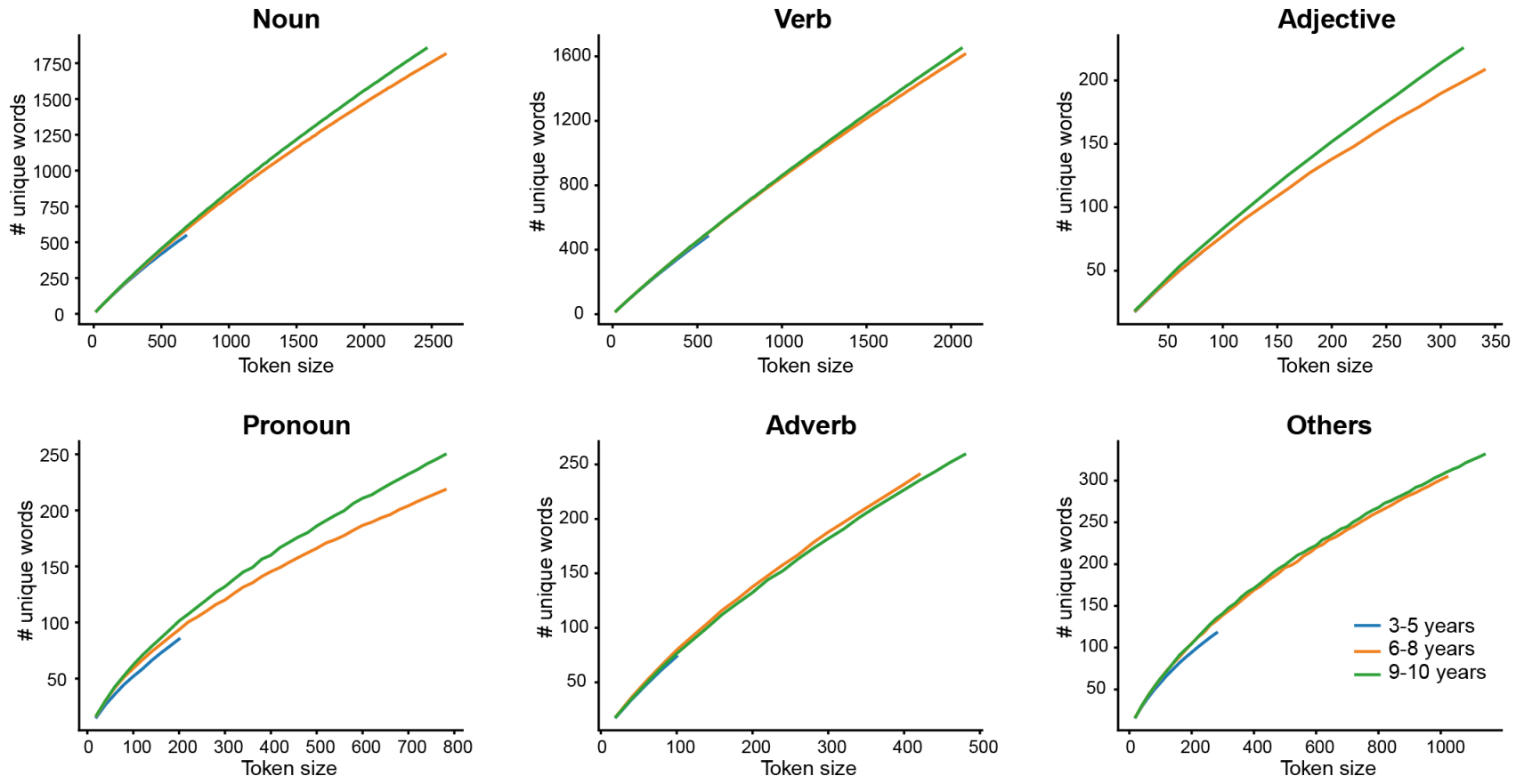
Expected trend:

Sensitive to corpora size

Increases in older books



The *manual* mapping of the psycholinguistic properties of such corpora



- **The number of unique words by parts of speech and book level.**

Examining small-sized corpora

- The selectively sampled corpus approximated the larger corpus for:
  - Occurrence statistics for a range of orthographic types irrespective of their frequency
  - Phonological and orthographic word lengths
- Longer corpora will by definition have more instances of unique units
  - Orthographic diversity therefore differs across bigger and smaller corpora

Expanding the evidence base

- Understudied languages may need to start with small-sized corpora.
- If a small-sized corpus must be used, then a non-sequential sampling protocol with shuffling of word lists makes the corpus statistics better approximate what is found in larger corpora.
- Such corpora may provide a developmental catalogue of real-world print encounters.

- This may support evidence-building on under-theorized aspects
  - Architectural principles of writing systems
  - Language features at the level of phonology, morphology and syntax

Acknowledgements

Aakash Agarwal, Sunila John, Shivani Tiwari and the TalkTogether Child-Directed Print Corpus team

Maggie Snowling, Shula Chiat, Carole Torgerson and the DFID Narrative Review team

Co-editors and contributing authors of the forthcoming volume on Global Variations in Literacy Development

Thank you!

Talk Together
Supporting Oral Language Development

Web: http://talktogether.web.ox.ac.uk
Twitter: @talktogproject