READ OXFORD

Book language

Promoting literacy and oracy in the early years via structured experience with written language

UNIVERSITY OF OXFORD

Nuffield Foundation

MEET THE TEAM

Professor Kate Nation: Principal Investigator, Department of Experimental Psychology, University of Oxford, and Dr Nicola Dawson: Research Associate, Department of Experimental Psychology, University of Oxford and Moor House Research & Training Institute, Moor House School & College.

COLLABORATORS

- Dr Yaling Hsiao: Formerly British Academy Post-Doctoral Research Fellow, Department of Experimental Psychology, University of Oxford (now Lecturer in Education, University of Birmingham)
- Dr Nilanjana Banerji, formerly Senior Children's Language Data Specialist, Children's Dictionaries and Learning at Home, Oxford University Press (now Language Information Specialist, Oxford Languages, Oxford University Press)
- Ms Yuzhen Dong, Doctoral Student, Department of Experimental Psychology University of Oxford
- Mr Alvin Tan, former undergraduate intern, now doctoral student at Stanford University
- Ms Sally Brockbank, former Research Assistant, Department of Experimental Psychology University of Oxford, now Speech and Language Therapist, John Radcliffe Hospital

STUDENT INTERNS AND UNDERGRADUATE PROJECT STUDENTS

Mohammed Amara
Jessica Bate
Bethany Biggs
Lily Cavey
Charlotte Dale
Rebecca Denison
Emma Doe
Emily Dowse
Songjun He
Carys Hoggan
Eleanor Holton
Emma Jackson
Joel Kovoor
Sean McCarron
Tsvetana Myagkova
Nilo Pedrazzini
Zoe Popescu
Talia Rabinowitz
Amelia Rock
Georgia Sandars
Jack Southall
William Thurwell
Catherine Wang
Rhianna Watt
Rebecca Williams

ADVISORY BOARD

- Nilanjana Banerji, Oxford University Press
- Dorothy Bishop, University of Oxford
- Megan Dixon, Education Consultant
- Charles Hulme, University of Oxford
- Kim Pickin, Story Museum, Oxford
- Maggie Snowling, University of Oxford

INTRODUCTION

By the time children start school, they have learned a huge amount about their native language. This includes how words sound and what they mean, and how they combine to represent events. The onset of literacy provides children with new opportunities. Once they can read and write, children can create and enter other worlds, real or imagined, and written language quickly becomes the major vehicle for knowledge acquisition, be it via formal education or reading for pleasure. The onset of literacy also brings challenges. Most obviously, children need to learn how to decode and identify individual words accurately and fluently, a journey that begins with the appreciation that (in languages like English) words contain letters, and that letters and letter patterns relate in systematic ways to how words sound. Alongside this, they need to understand what they have read. Reading comprehension is multifaceted and complex, drawing heavily on the knowledge and skills children have already built through speaking, listening, and communicating from birth onwards. There is scientific consensus that oral language provides the critical foundation for learning to read. That written language is predicated on spoken language makes sense given our evolutionary history – spoken language is biologically primary whereas writing systems are cultural inventions – and our developmental history, where children have typically had considerable experience with spoken language before they come to the task of learning to read.

From this perspective, we can consider the early stages of learning to read as one in which children must learn how their orthographic system works so that they can identify written words, and from this access their spoken language repertoire to construct meaning from text. This characterisation is formalised in the Simple View framework that sees reading comprehension as the product of word reading and listening comprehension. The simple view has considerable empirical support: when each component is measured appropriately, they together account for substantial variation in reading comprehension. It also features in education policy and practice, reflecting its utility for the classroom. Despite its many strengths, there are limitations to the simple view framework (e.g., Catts, 2018; Nation, 2019). These do not detract from its importance or its central message, but they do serve to highlight the complexity of reading.

One complexity concerns the language of books itself. Text is not speech written down. Spoken language is generally in the moment and in the context of social interaction. Text, by contrast, is remote. Written language cannot rely on situational cues or shared context to deliver its message. Instead, the text itself must provide sufficient specification so that the reader can infer the meaning intended by the writer. This means that written and spoken language differ in important ways. Numerous studies of adult language have shown core differences in tone and formality. There are also substantial differences in linguistic complexity: vocabulary is denser and more diverse in writing and some syntactic structures are rare in conversations but common in writing (e.g., Biber, 1988; Roland et al., 2007). This increased complexity allows writing to represent meaning in a way that is decontextualized until it is re-constructed by the reader. Thus, while oral language is a necessary foundation for learning to read, exposure to conversational language is not sufficient. Children also need to learn 'the language of the book'.

There is evidence that opportunities for this learning start early. For example, Montag et al. (2015) compared the language of children's picture books with the language experienced via child-directed speech. Books contained more words, and a greater variety of words. This suggests that well before children can read themselves, they can experience book language via shared reading. This is an important finding as there are large differences in the quality and quantity of shared book reading at school entry (e.g., Mol et al., 2008). These differences are strongly associated with differences in the home literacy environment and social advantage[11] meaning that some children are at a serious disadvantage when it comes to learning to read. This observation supports an intervention approach designed to increase shared reading between caregivers and their pre-schoolers. While some positive effects have been reported, especially in terms of increasing reading enjoyment and following intensive intervention, transfer to improved language as assessed by standardised tests is slight in hard-to-reach populations and in shorter-term studies (e.g., Lingwood et al., 2020; Noble et al., 2019).

AIMS OF THIS PROJECT

Our purpose was twofold. First, we aimed to systematically define, quantify and review the nature of children's book language. This allowed us to identify 'literary' vocabulary and syntax

– language that children are more likely to learn from books than from everyday conversations. Second, we systematically introduced children to book language via shared reading and monitored the impact of this through a carefully controlled experiment with 4-7 year-old children. Our intention was to discover whether structured book language experience provides key and specific support for learning in the early years. Meeting these two aims required two very different methodological approaches, as detailed in the next two sections of this report. By addressing these two aims, we hoped to be well-placed to consider the feasibility, scalability, and utility of systematic exposure to book language as an approach to secure the foundation for learning in the curriculum, especially for those starting school from a position of disadvantage.

PART I: UNDERSTANDING THE NATURE OF CHILDREN'S BOOK LANGUAGE

RATIONALE AND BACKGROUND

Corpus linguistics involves computer-based analyses of language use across large collections of naturally occurring datasets (i.e., language corpora). Inspired by previous work with relatively small corpora comprising child-directed speech and child-directed text (Montag et al., 2015; Montag & MacDonald, 2015), we decided to undertake a series of developmental cross-corpus analyses. This allowed us to harness the power of big data to quantify and specify the nature and content of book language across age and to detail how it differs from spoken language and varies by genre.

METHOD

Table 1 summarises the corpora used in our study. We extracted child-directed speech from UK samples in the Child Language Data Exchange System (CHILDES; childes.talkbank.org). This provided a proxy for language input via conversations children hear in the home, in the pre-school years. The Oxford Children's Corpus (OCC) is the largest corpus of materials written for children in the English-speaking world. The corpus was created by and for Oxford University

Press (an academic department of Oxford University). It is a dynamic and growing corpus, and we were fortunate to have access and to collaborate with colleagues at OUP. It is in two main parts – children's books (giving a proxy for what children read) and children's own writing, comprising stories submitted by children from across the UK for 500 Words, an annual children's writing competition hosted by BBC Radio 2 in collaboration with OUP.

As we were mainly interested in language input in this project, this report focuses on comparisons between child-directed speech and children's books. As some of our published papers report analyses that also add in comparisons across children's own writing, we note the writing corpus here, but do not discuss those findings in any detail.

We initially envisaged (in the grant proposal) using the OCC-Reading to answer our research questions. However, when we started the work, we realised that the corpus did not include many books targeted at pre-schoolers. As we were interested in language input prior to school entry, we decided to build our own corpus, the ReadOxford corpus, to provide the resource needed. The creation of this corpus is described by Dawson et al. (2021) and our materials have been made available to other researchers.

Analyses were conducted in R and python. As some analyses were quite complex, we refer readers to our published papers for full details and links to analysis pipelines and scripts (and links to project outputs are provided later in this report).

Table 1. Overview of corpora analysed in this project

Speech CHILDES	Picture Books ReadOxford	Reading Books OCC (OUP)	Children's Writing OCC (OUP)
0-6 years 10 corpora from English-UK database	0-6 years 160 books, mainly fiction	~21,000 documents Meta-data on targeted Key Stage 1-4 and genre	~ 1 million stories from 5-13 year-olds, BBC Radio 2 500 Words competition

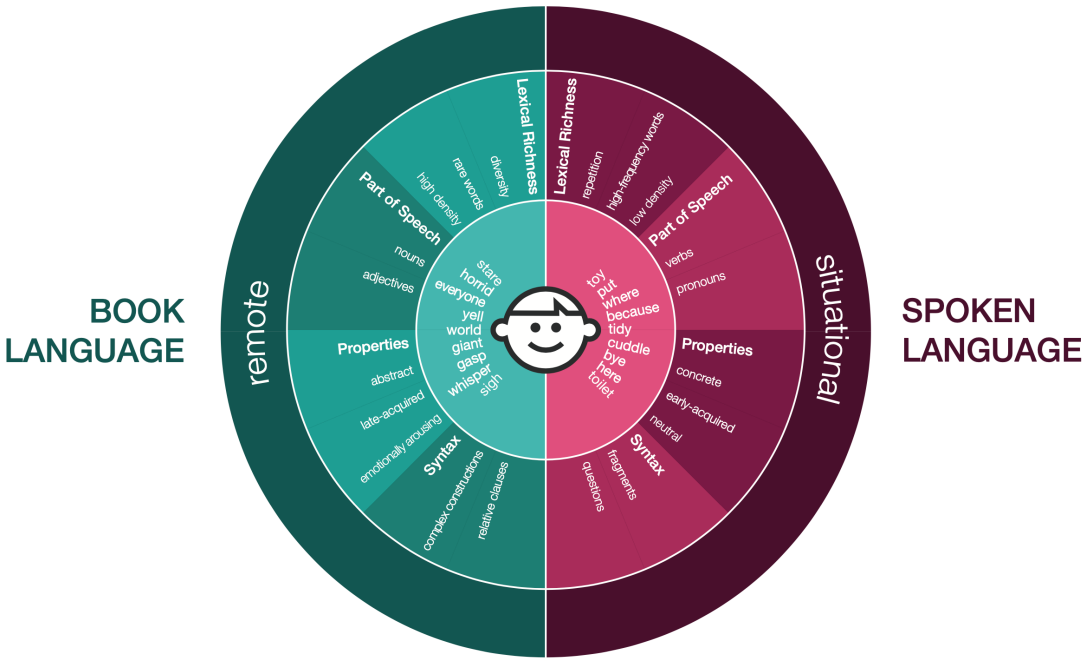
~3.8 million words	~320,000 words	~47 million words	~ 1 million stories (each of 500 words)
Caregiver-child conversations in the home	Published picture books for shared reading	Mainly fiction, some non-fiction, from books, websites, magazines	Any topic, only constraint is length (max 500 words)

KEY FINDINGS

We have published an accessible and open access review paper (Nation et al., 2022) that describes some of our findings to date, and those of other researchers. The infographic in Figure 1 summaries some of the key differences between book language and conversational spoken language. Note, while making a distinction between book language and spoken language, it is important to remember that factors such as formality and genre influence patterns of language use by adults within each modality, and social media reminds us that both written and spoken language continue to adapt and evolve. These factors are likely to be evident in child-targeted language too, reflecting sensitivities to discourse patterns within and across modalities rather than a clear dichotomy between speech and writing.

We expand on the infographic by considering findings across three sets of analyses focusing on lexical, syntactic, and morphological factors.

Figure 1. Infographic summarising key differences in the quantity and quality of language input children experience via book language and spoken language (for data and annotated examples, see Dawson et al., 2021; Hsiao et al., 2022)



(I) LEXICAL RICHNESS

Dawson et al. (2021) investigated lexical richness in the ReadOxford picture book corpus in comparison to child-directed speech. Richness was quantified in three ways, namely diversity, density and sophistication. Lexical diversity provides an indication of vocabulary breadth and is usually measured using type-token ratios. Token frequency refers to the total number of words whereas type frequency counts the number of unique words. From Figure 2 we can see that there are more types at any token sample size in children's picture books, indicating that they contain more diverse vocabulary than child-directed speech. Lexical density captures the proportion of lexical items (e.g., nouns, lexical verbs, adjectives, and adverbs derived from adjectives) in a language sample, relative to the total number of words. As shown in Figure 3, we found that children's books contain significantly more content words than child-directed speech, consistent with books being more information dense.

Figure 2. Mean number of word types at different sized samples of word tokens randomly selected from the picture book and spoken language corpora.

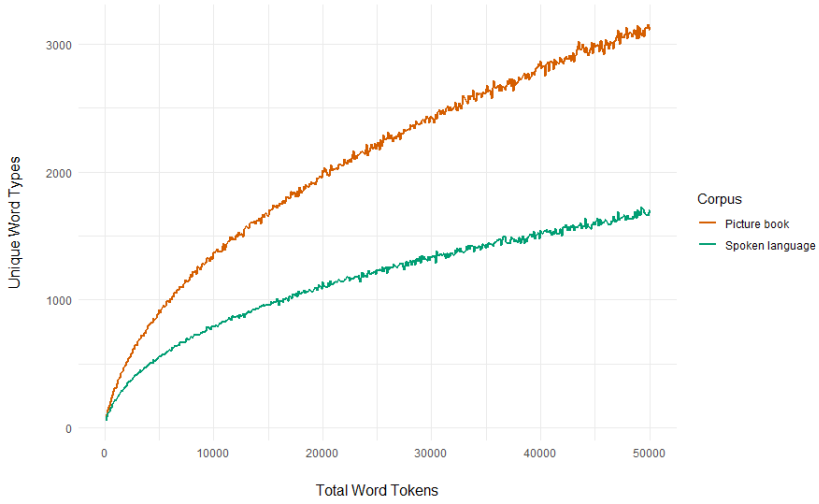
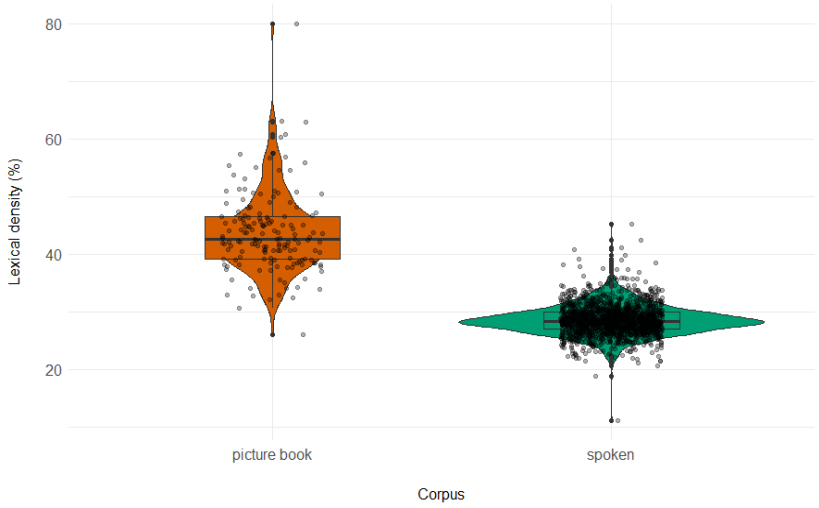


Figure 3. Percentage lexical density across picture book and spoken language corpora, plotted by individual document (picture book corpus) and conversation (spoken language corpus)



Like lexical density, measures of lexical sophistication shed light on the types of words contained within a language sample, and in particular, whether those words are skewed towards one end of the frequency distribution. Our analyses found that book language words were rarer than conversation words, that is, they were less likely to be in the top 1000 most common words in English. This tells us that access to picture books increases the likelihood that children will experience rarer word types that they would not otherwise encounter through conversation alone. We also found that book words were longer and more morphologically complex.

Follow-up analyses used keyword analysis to identify those words that were most uniquely representative of books and those least uniquely representative. Overall, book words were later acquired, more abstract, and more emotionally arousing than the words more common in child-directed speech. Table 2 provides some examples that clearly illustrate these conclusions.

Table 2. Example words from the keyword analyses illustrating the nature of 'book words'

Example words **most** representative of book language			Example words **least** representative of book language		
Stare	Deep	Shriek	Yeah	Mmmm	Oy
Voice	Gasp	Mutter	Alright	Careful	Car
Begin	Whisper	Large	Darling	Shall	Yesterday
Horrid	Dad	Cheer	Pardon	Poorly	Naughty
Suddenly	Leap	Shout	Okay	Nursery	Yum
Father	Sigh	Dream	Whoops	Yes	What
Everyone	Perfect	Each	Ok	Penguin	Today
Yell	Enormous	Towards	Hmm	Want	We
Giant	Thought	Silence	Wee	You	Doll

In summary, listening to book language provides exposure to vocabulary that is quantitatively and qualitatively different to that experienced via day-to-day conversation. We discussed

these findings in the broader literature in our paper. We also speculated that given words in books are more advanced, the impact of variation in exposure to book language may relate more closely to the skills that underpin children's emerging literacy. The words that children encounter in picture books are by definition more characteristic of the literary domain. Importantly, experience is key: exposure to picture books via shared reading allows children to start encoding the phonological forms and meanings of more advanced words across different contexts from an early age. Over time, this experience will shape language development and provide a strong foundation to literacy. While there are many potential benefits of shared reading for children's development, our findings suggest that one of the key contributions may stem from the language of the books themselves, and specifically the rich and diverse lexical input they offer. We begin to test this hypothesis in the research described in Part II of this report.

(II) SYNTACTIC COMPLEXITY

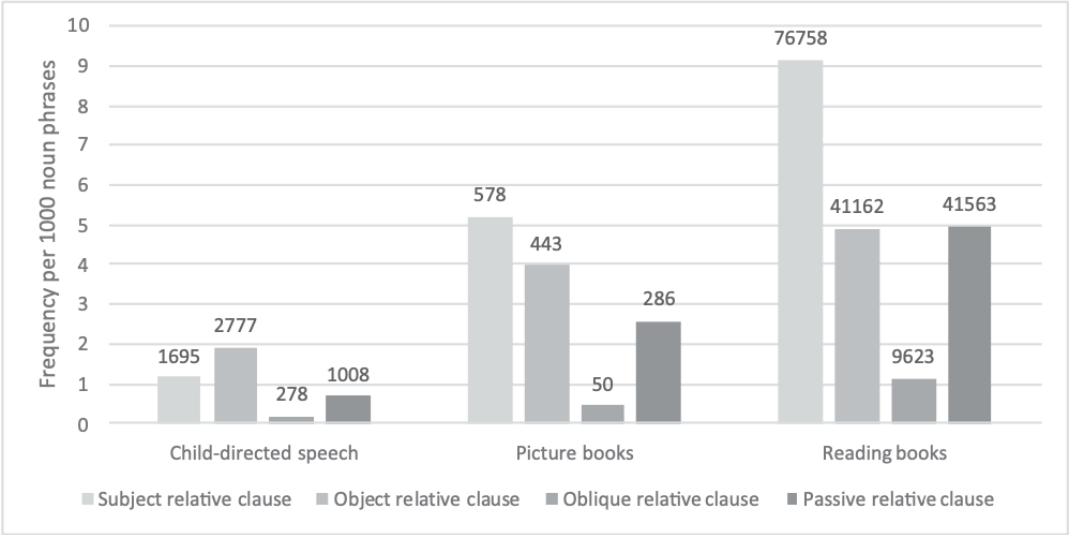
Words are experienced in sentences and syntactic structures, and there is previous work showing that children's books tend to contain more complex syntactic structures than child-directed speech. For example, Cameron-Faulkner and Noble (2013) analysed 20 picture books written for 2-year-olds and found they contained more complex constructions than child-directed speech. Montag and MacDonald (2015) also reported more complex syntax (e.g., object and subject relative clause, passive) in their analyses of 97 texts targeted at 4-16 year-olds; there was also a positive association between amount of complex syntax in book language and intended age. Importantly however, their corpus contained materials intended to be read by the child, rather than read to the child. This would naturally constrain the number and complexity of texts aimed at younger children, given constraints on decoding in early reading development. We aimed to build from these findings on relative clauses and investigate whether they extend to children's picture books, using the ReadOxford corpus as well the OCC-Reading corpus (Table 1), and comparing both with child-directed speech. Note, relative clauses are the part of a sentence connected to its main part by a word such as *that, who, which*; e.g., She loved the garden *that she used to tend*.

Our paper (Hsiao et al., 2022) is complex and detailed, and we refer the reader to the full text where the methodology and findings are described in full. In brief however, we automatically

identified and extracted relative clauses from the three corpora, and classified them into four linguistic categories (subject, object, oblique and passive). Given there are some concerns with automatic coding, we manually checked 1000 random exemplars drawn from each corpus. This established that although not perfect, our automated methods were largely accurate and valid.

As is clear from Figure 4, all types of relative clauses were less frequent in child-directed speech than in either sample of book language. The contrast between picture books and child-directed speech is particularly informative as both contain language targeted primarily at pre-schoolers. Even when the age of the child is comparable, there were more relative clauses in book language than spoken language (12.24 vs. 3.97 relative clauses per 1000 noun phrases). Across the two types of book language, picture books contained fewer relative clauses than books written for children to read themselves (12.24 vs. 20.18 relative clauses per 1000 noun phrases). The pattern of relative frequency across the four different types, however, was similar between the two book language corpora, with subject relatives most common. In child-directed speech, object relatives were most frequent. In all three corpora, oblique relative clauses were the rarest among all types.

Figure 4. The frequency distribution of the four types of relative clause per 1000 noun phrases across corpora. Raw frequency is shown as labels.



We also took advantage of the meta-data in the OCC-Reading corpus to take developmental slices through the corpus by targeted age (as approximated by Key Stage) and by genre (comparing fiction and non-fiction). We saw differences in the type and distribution of relative clause across these corpora and sub-corpora. Object relatives were the most common type in child-directed speech but were less common in book language, and in nonfiction in particular. In contrast, passive relatives were rare in child-directed speech but became gradually more common in texts for older children, and in nonfiction. Subject relative clauses occurred more often in picture books for pre-schoolers than speech directed to children of similar age; they were more frequent still in books for independent reading, and in nonfiction. Although oblique relative clauses were the rarest type across all corpora, they were more common in books than in speech, and in fiction than nonfiction. Taken together, these frequency counts and cross-corpus comparisons show that book language provides children with exposure to variations in complex grammar from the outset and as targeted developmental level of text increases, so too does the amount and nature of complex grammar.

If children experience more complex syntax as a function of their exposure to written language, we would predict that this experience and input should be reflected in the types of sentences that children or adults find easier or more difficult to comprehend. We were not able to look at this directly in our own work, but in our discussion in Hsiao et al., we made systematic connections between the lexical syntactic properties of the different sentence types (e.g., noun animacy, verb transitivity, pronoun vs full noun status) and how readily these are understood, according to sentence processing literature (largely from adults). We saw synergies between the patterns we identified in input distributions (i.e., frequency in book language) and ease of processing (as reported in the experimental literature).

Much like our work on lexical richness, this work highlights the clear need to investigate differences between spoken and written language targeted at children. Books, even those written for pre- schoolers to hear in the context of shared reading, contain more relative clauses than child-directed speech. Our findings replicate and extend previous smaller scale studies and show that as sophistication of text grows with increases in targeted age, so too does the frequency of relative clause usage. Importantly, it is not just the number of relative clauses that changes but also their type and distribution: both picture books and reading

books are dominated by subject relative clauses, different from speech, and book language contains dramatically more obliques and passives than child- directed speech. These changes are evident even in the youngest developmental slice through the reading corpus, capturing books written for 5-7 year-old children. There are also differences by genre with subject and passive relative clauses being more common in nonfiction.

(III) MORPHOLOGICAL COMPLEXITY

Our work on lexical richness (Dawson et al., 2021) indicated that book words in the children's picture books were more likely to be morphologically complex than those more frequent in child-directed speech. This is an interesting and potentially important finding, given discussion and debate as to when and how children become sensitive to morphological regularities through reading. This sensitivity might itself be a marker of the transition to more expert reading, once the basics are in place (e.g., Rastle, 2019). To inform these discussions, we tracked children's experiences of written morphology by analysing the OCC-Reading (Table 1), a large corpus of children's reading materials spanning a target age range from 5 to 14 years. We (Dawson et al., 2023) examined frequency distributions of morphologically complex words by target age and genre, as well as type and token frequencies for 80 individual derivational suffixes. We found that the proportion of morphologically complex words – and derived words particularly – increased in line with target age, and that nonfiction contained more complex words than fiction. Frequencies of individual suffixes also varied by target age and genre, with Germanic forms more common in fiction and texts for younger children, and Latinate forms more common in nonfiction and texts for older children. These findings provide a comprehensive picture of how children's experience with written morphology changes over the course of reading development. As with our earlier work, these findings also invite us to speculate on different learning opportunities provided as a function of reading experience. Children who can read, and children who read broadly, will experience more words and more complex words than those who read less. We proposed that this experience will shape children's reading system such that reading becomes more expert, as revealed by increasing sensitivity to morphological complexity during word recognition itself (for review, see Rastle, 2022). While this is not of central focus here, we refer interested readers to our paper (Dawson et al., 2023) where we discuss our corpus findings in the context

of developmental changes in morphological processing. We also reflect on the benefits and limitations of using large-scale natural language datasets more generally.

PART II: UNDERSTANDING WHAT CHILDREN LEARN FROM BOOK LANGUAGE EXPERIENCE

RATIONALE AND BACKGROUND

As discussed above, book language provides a very different kind of input to spoken conversation, with more varied, sophisticated and abstract vocabulary, and rarer and more complex sentence structures (Dawson et al., 2021; Hsiao et al., 2022; Montag et al., 2015; Nation et al., 2022). These differences matter because children learn from the language they hear (Aslin & Newport, 2014). Once children can read, they can independently access the rich and diverse language that books offer, but until then, parents and caregivers reading *to* children provides an important pathway to this input early in development. Yet we know that access to shared reading activities in the early years varies hugely from child-to-child, and this variation has been linked to language and literacy development (Bus et al., 1995; Hamilton et al., 2016; Mol et al., 2008). This strand of the project asked whether systematic exposure to book language directly benefits children's language, focusing on the diversity and sophistication of words in books that emerged as key features in our corpus analyses.

Using keyword analysis, we identified words that were most representative of children's books compared to child-directed speech. 'Book words' included several sets of synonyms or near-synonyms (e.g., *shriek*, *yell* and *shout* or *whisper* and *mutter*).

Examples of words most representative of books		
stare	deep	**shriek**
voice	gasp	**mutter**
begin	**whisper**	large
horrid	dad	cheer
suddenly	leap	**shout**
father	sigh	dream
everyone	perfect	each
yell	enormous	towards
world	reply	cave
giant	thought	silence

These sets of words represent nuanced ways of capturing the same underlying concept. Variability in language input might support language learning and processing in both children and adults (Cassani et al., 2018; Hadley et al., 2016, 2017; Hills et al., 2010; Hoff & Naigles, 2002; Tamminen et al., 2015) and lexical diversity specifically has been linked with rate of vocabulary growth in children (Hsu et al., 2017; Rowe, 2012). If books provide encounters with different words that share a core meaning (e.g., synonyms), this may support children's word knowledge in at least two ways: firstly, by exposing them to a broader range of word types, many of which occur infrequently in English, and secondly, by facilitating associations between related word meanings. Our aim was to investigate this hypothesis by manipulating

lexical richness, which we defined as both lexical diversity (the number of unique word types in a language sample) and lexical sophistication (the proportion of rare word types in a language sample), in stories and examining the effect on children's word knowledge and narrative production.

REGISTERED REPORT

The study was submitted as a registered report. This meant that the proposed rationale, methods and analysis plan were submitted to a journal and peer reviewed prior to the start of data collection. The benefits of this route to publication (see Figure 5) are that reviewers provide feedback on the study design while there is scope to make changes, and acceptance for publication is based on the quality of the proposed methods rather than the outcomes of the study. This approach has wider advantages in reducing publication bias, where positive results are more likely to be published than negative or inconclusive findings. In-principle acceptance is awarded following successful review of a Stage 1 report outlining the background literature, methods and proposed analysis plan. Full acceptance is given following successful review of the full manuscript once the study has been run, provided authors have followed the protocol outlined in the Stage 1 report.

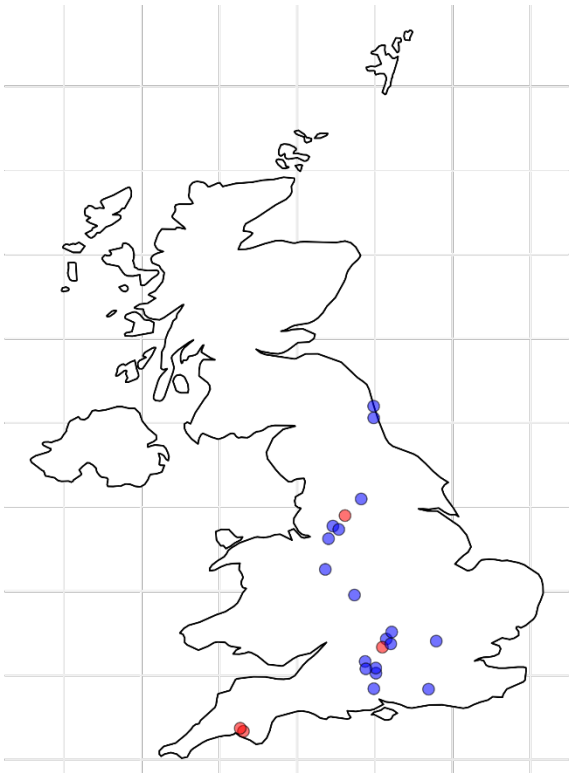
Figure 5. Infographic summarising the Registered Report, re-printed from the Open Science Framework (https://help.osf.io/article/159-submit-a-registered-report)

This study currently has in-principle acceptance as a Stage 1 registered report in *Journal of Child Language*. The full Stage 1 report and associated materials (e.g., power analysis) are available at: https://osf.io/eqdtv.

METHOD

PARTICIPANTS AND SETTINGS

The study was conducted over multiple sessions in mainstream school and nursery settings within the UK. A total of 180 children participated in the project - 153 attending Years 1, 2 or 3 in school and a further 27 based in nurseries or early years settings. All children were aged between 4-7 years.



Approximate locations of participating settings: schools indicated in blue and nurseries in red

BASELINE LANGUAGE MEASURES

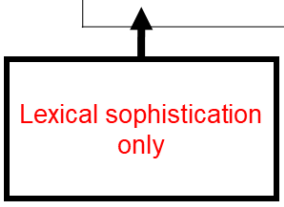
Children's baseline language ability was measured using the Language Screen app, developed and standardized as part of the Nuffield Early Language Intervention (NELI) programme (West et al., 2021). This test assesses four core language skills: expressive vocabulary, receptive vocabulary, sentence repetition, and listening comprehension. We also assessed baseline narrative ability using the 'narrative recall' subtest of the Assessment of Comprehension and Expression 6-11 (ACE 6-11; Adams et al., 2001) in which children listen to and retell a short pre-recorded story with the aid of illustrations.

STORIES

We created eight colour-illustrated stories designed to be read to young children. All text was removed from the books so that children's reading ability did not influence what they learned from the stories. There were three versions of each story. The **lexically-rich** version included a set of four verb synonyms (e.g., *laughed, chuckled, giggled, guffawed*) and a target set of four adjective synonyms (e.g., *sad*, *miserable*, *glum*, *dejected*).

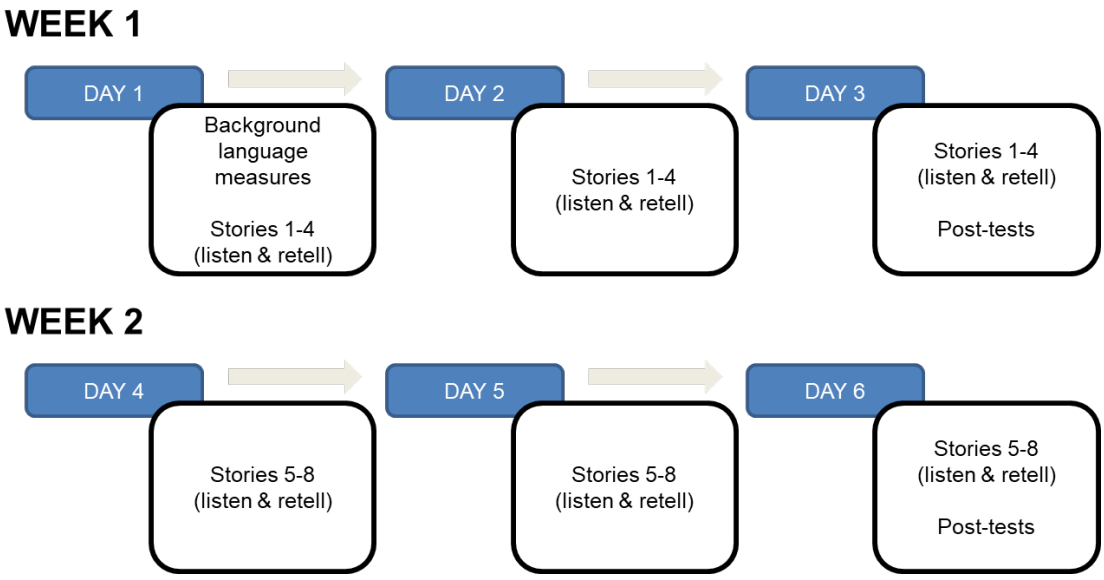
Each set of synonyms had a base word (*laughed*, *sad*), which was the word acquired earliest in development and which occurred commonly in English. The synonyms also included an advanced word (*guffawed*, *dejected*), which was the word that was acquired latest in development and which occurred relatively infrequently in English. The **control-advanced** version of the stories were identical except that the advanced word in each set (*guffawed*, *dejected*) was repeated four times in place of the synonyms. The **control-base** version repeated the base words four times (*laughed*, *sad*). Each child was randomly allocated to one of the three different conditions. The table below outlines how lexical diversity and sophistication varied across the three conditions:

	Target **adjectives** (x4)	Target **verbs** (x4)
Lexically rich	sad, miserable, glum, dejected	laughed, chuckled, giggled, guffawed
Control base	sad, sad, sad, sad	laughed, laughed, laughed, laughed
Control advanced	dejected, dejected, dejected, dejected	guffawed, guffawed, guffawed, guffawed



READING SESSIONS

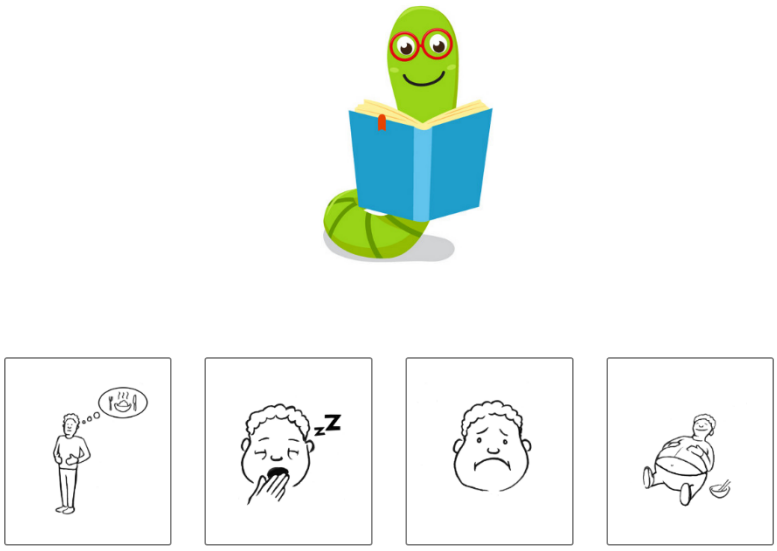
The eight stories were split into two sets (A and B), with four stories in each. Children were read the first set of four stories on three separate days over the course of one week. In each session, children heard each story read aloud, and were then immediately asked to retell that story with the aid of the illustrations. In the first session, children also completed the background language measures, and in the third session, they completed the post-tests. They then repeated the process with the other set of stories.



POST-TEST MEASURES

Children's knowledge of the advanced words and semantic associations between the synonyms was measured using two post-tests. These were administered on a tablet following the third reading session for each set of stories.

In the tasks, children were introduced to a character called 'Billy the Bookworm' who enjoyed learning new words but needed some help. In the first task, children heard the advanced word from each synonym set (e.g., *dejected*, *guffawed*) in isolation and were asked to click on the picture that best showed the meaning of that word.



"Click on the picture that best shows the meaning of dejected"

In the second post-test, children helped Billy the Bookworm sort pairs of words into either a blue book (for word pairs that had the same meaning) or an orange book (for word pairs that had different meanings). Children would hear the base verb or adjective from each set paired with a) the associated advanced synonym, b) an associated intermediate synonym, or c) an adjective or verb from a different set of synonyms. Children then selected the blue or orange book depending on whether they classified the word pairs as 'same' or 'different'.

KEY FINDINGS

Do children who experience lexically richer stories produce a lexically richer output when asked to retell those stories, and does this effect increase in line with number of exposures?

We looked at the language children used to retell the stories in each of the three sessions. In this measure we examined lexical diversity of children's story retellings, and whether this differed depending on the version of the stories they heard and the number of times they had heard them. Lexical diversity was calculated using a measure called 'Moving Average Type-Token Ratio (MATTR)', which computes the number of unique words in a language sample, taking account of the total number of words in that sample. Results below are based on a subsample of recordings from 36 children:

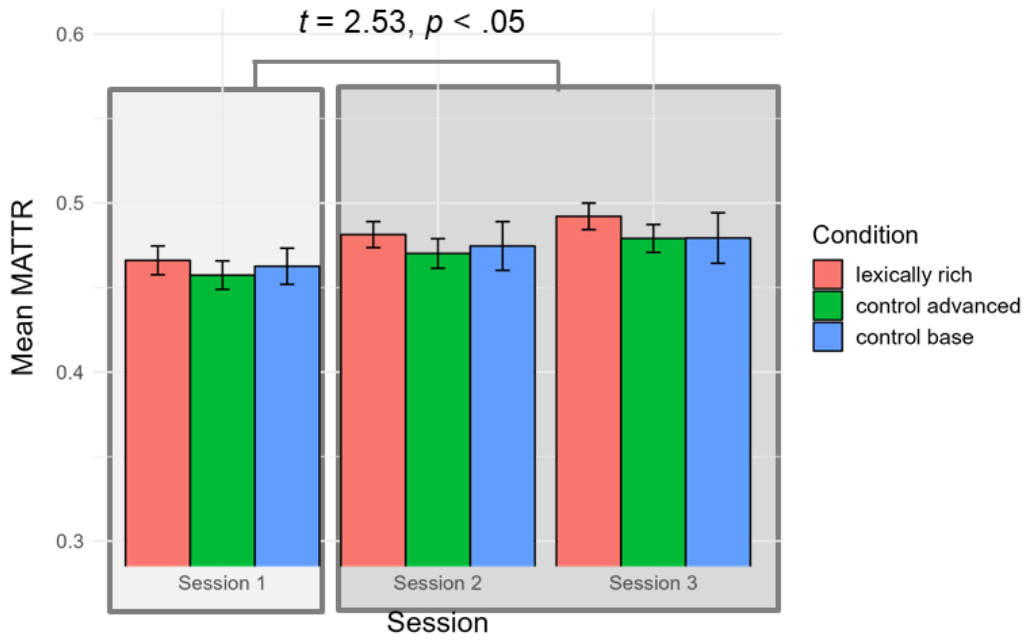


Figure 6. Mean lexical diversity in children's story retellings by condition and session

We found that lexical diversity increased the more times children heard and retold the stories (i.e., they used a broader range of word types in sessions 2 and 3 compared to session 1). However, the version of the stories that children heard did not affect the range of words they used. It is important to note that this effect may be difficult to detect with the current sample of 36; the complete set of data will be reported in the Stage 2 Registered Report, and the publication will be available Open Access and on our website.

Is there evidence that children better understand the meanings of more sophisticated word types when they repeatedly encounter them in stories? Is frequency or diversity more important?

We examined children's understanding of the advanced word (e.g., *dejected, guffawed*) in each set of synonyms based on their performance on the first post-test task. Our question was whether it was more helpful for children to hear the advanced words in stories multiple times (i.e. the control advanced condition), or whether it was better to hear them in combination with more basic words sharing the same core meaning (i.e. the lexically rich condition). This analysis was based on responses from the full sample of children (N=180).

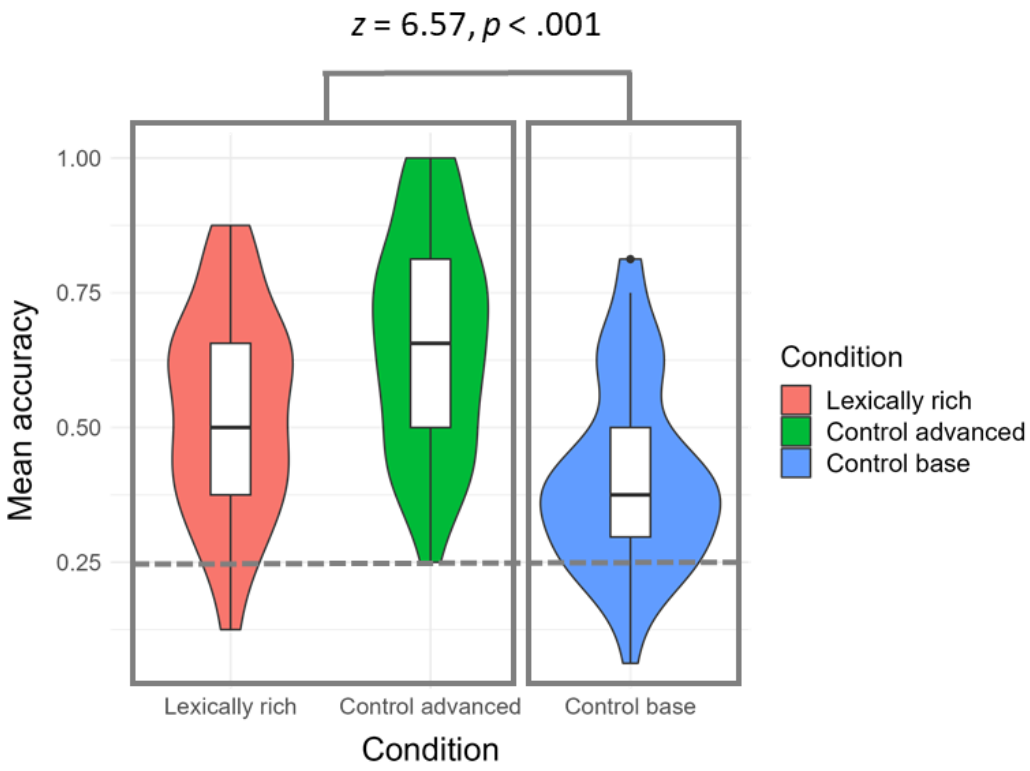


Figure 7. Mean accuracy on comprehension of advanced words testing the effect of exposure

There was evidence that children benefitted from encountering the advanced words in stories: they were more likely to choose the picture representing the appropriate meaning of the word if they had heard the stories in the lexically rich and control advanced conditions compared to the control base condition. A mean accuracy score of 0.25 represents chance level because the multiple choice had four options. The plot indicates that children had some prior knowledge of the advanced words, but performance was still improved if they encountered them in the stories.

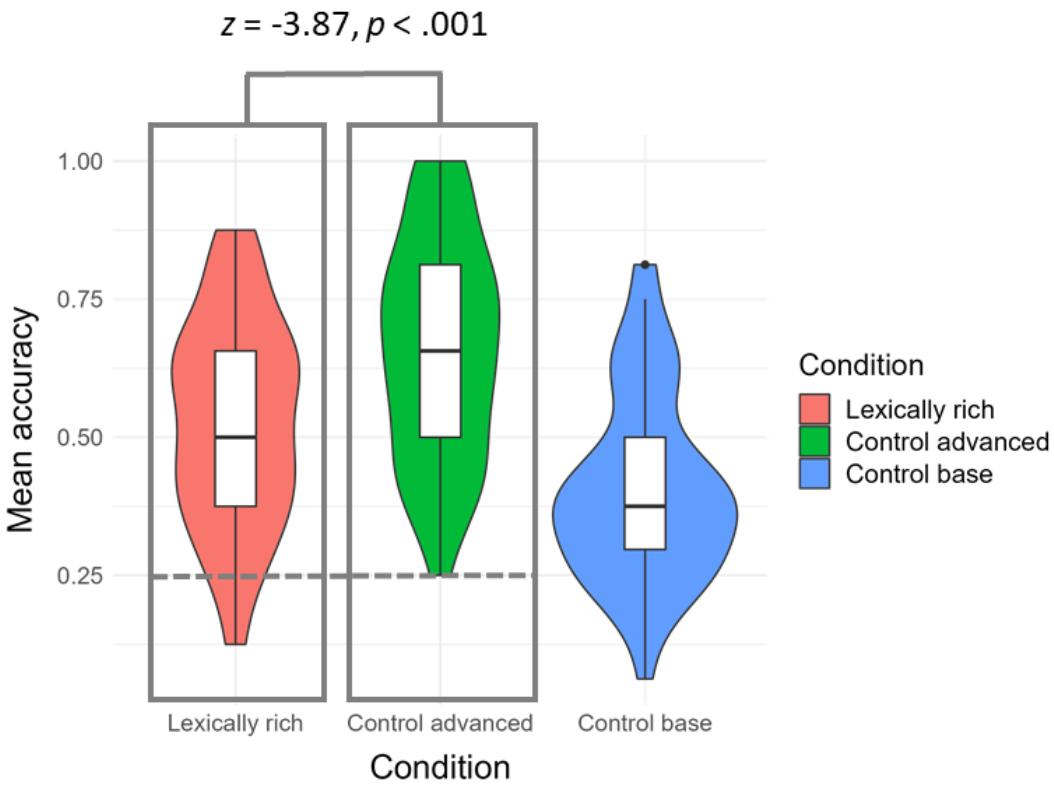


Figure 8. Mean accuracy on comprehension of advanced words testing the effect of frequency vs. diversity

Our second question was whether it mattered more to hear the advanced words multiple times, or to hear them in combination with more basic synonyms. The plot above indicates that children performed better on the multiple-choice post-test when they heard the advanced words multiple times in the stories.

Are children who are exposed to multiple words with similar meanings in stories better at recognising the semantic association between those words?

We evaluated children's understanding of how words within the synonym sets were related in meaning using their performance on the second post-test. In this task, children responded to pairings of both the base word and advanced word (e.g., *sad – dejected*) and the base word and one of the intermediate synonyms (e.g., *sad – glum*). We anticipated that children in the lexically rich condition would perform best for base-intermediate pairings (e.g., *sad – glum*), while children in the control advanced condition would perform best for base-advanced pairings (*sad – dejected*). There was no difference in how well children judged the relationships between the base word and its synonyms between children in the lexically rich

condition and children in the other two conditions (see Figure 9). However, children in the control advanced condition performed better overall than children in the control base condition (see Figure 10), providing further evidence that hearing the advanced words multiple times in stories led to better understanding of their meanings.

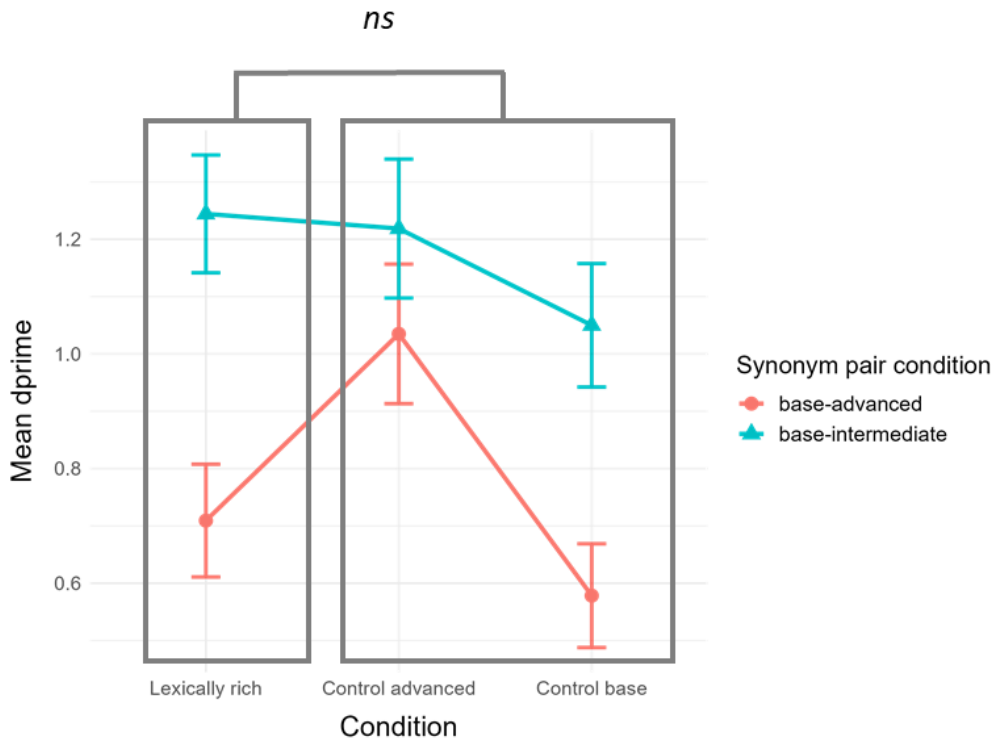


Figure 9. Performance on task judging semantic associations between synonyms comparing lexically rich condition to the average performance across the control conditions

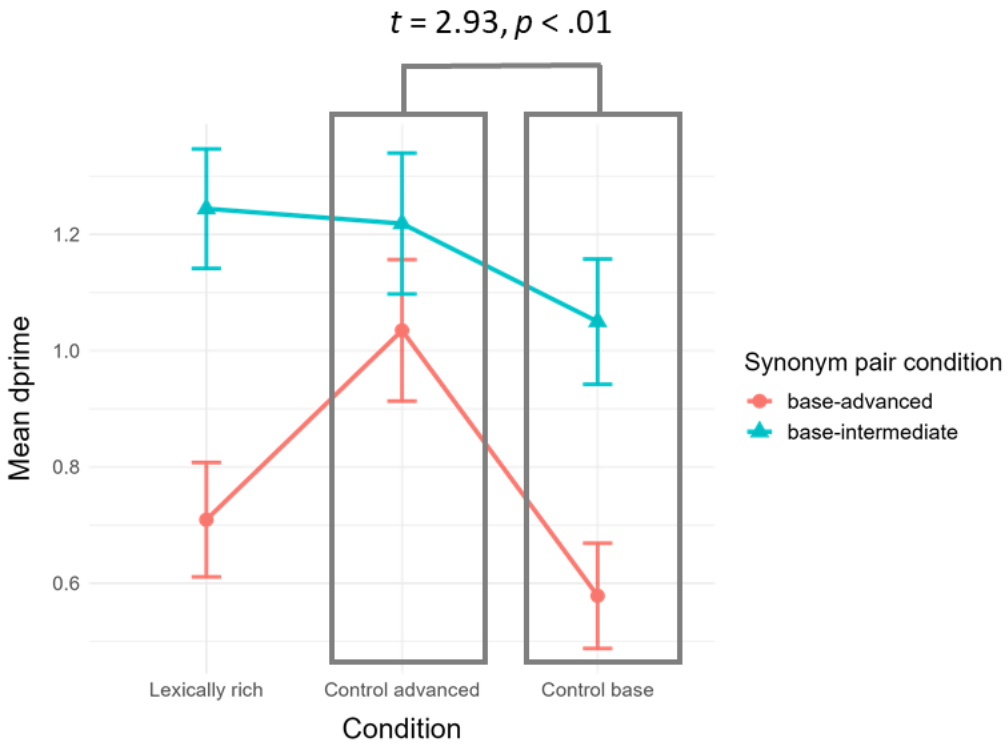
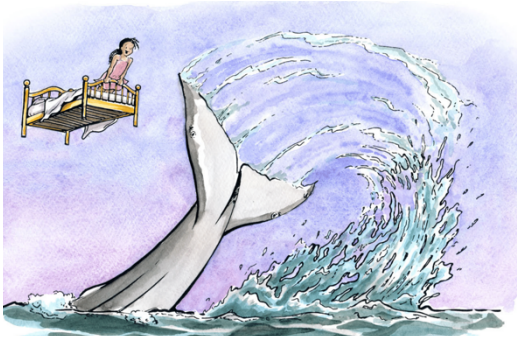


Figure 10. Performance on task judging semantic associations between synonyms comparing control advanced condition to control base condition

SUMMARY

This study provides evidence that children learn about language through hearing illustrated stories read aloud. Firstly, we found a benefit to re-reading stories to children. Children use more diverse language themselves when they retell stories that they have heard before. One possibility is that as children become more familiar with the macrostructure of the stories (i.e. the plot and sequencing of events), they are able to focus more on the microstructure (e.g., word choice), leading to a broader range of vocabulary in their retellings. While there is currently no evidence that children used more diverse language when they heard more diverse language in the input, this effect is likely to be small and may emerge with a larger dataset.



Secondly, we showed that children learn something about the meanings of sophisticated words after hearing them in stories on three occasions, even if they only appeared once. However, they showed

better understanding of their meanings if they encountered them four times compared to one. Because stories typically contain more sophisticated words than speech, reading stories to children may provide a rich source of advanced language input for children to learn from.

Finally, when children encounter advanced words in stories, they are able to associate them with known words spontaneously: there does not seem to be an added advantage to experiencing both words in the same story context. Nevertheless, encountering a diverse range of vocabulary through stories does introduce children to words that they may not otherwise encounter in everyday speech.

CONCLUSIONS

Learning to read is a key milestone in children's development, yet the language that children must master to become a fluent reader is starkly different to the language they hear in everyday conversation. This project systematically identified ways in which the language of children's books diverges from child-directed speech, and how features of written language pattern differently according to genre and target age. Our findings indicate that books offer a unique contribution in relation to the range and types of words that children encounter, and the complexity of syntactic structures. Crucially, these differences emerge in texts written for children who cannot yet read themselves, such that shared reading offers opportunities for very young children to learn from this input and build a foundation for later literacy. Our experimental work shows that children respond to the language that they hear in a shared reading context, using richer language themselves when they hear stories multiple times, and learning about the meanings of advanced words that they would rarely encounter in conversation. These findings build on shared reading intervention approaches, and they point to a pivotal role for the language of the books themselves in providing key and specific support for learning in the early years.

> *"The books transported her into new worlds and introduced her to amazing people who lived exciting lives. She went to Africa with Ernest Hemingway and to India with Rudyard Kipling. She travelled all over the world while sitting in her little room in an English village"*
>
> *- Roald Dahl, Matilda*

PROJECT OUTPUTS

Our papers are Open Access. Each paper contains links to repositories where we share data, analysis plans and code. It is not always possible for us to share all data (e.g., the Oxford Children's Corpus in copyrighted and commercially sensitive) but we aim to provide detailed information so that our pipelines and analysis protocols clear and transparent.

** The corpus research described in Part I of this report has contributed in important ways to other projects. We gratefully acknowledge the support of the Nuffield Foundation in all of this related work (see outputs marked **)

PUBLISHED/ACCEPTED IN-PRINCIPLE

Dawson, N.J., Hsiao, Y., Tan, A.W.M., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*, DOI: 10.34842/5we1-yk94

Dawson, N.J., Brockbank, S., & Nation, K. (2021). Effects of lexical richness in storybooks on children's narrative retellings and word knowledge. Stage 1 Registered Report, accepted in-principle, *Journal of Child Language*, https://osf.io/eqdtv

Dawson, N.J., Hsiao, Y., Tan, A.W.M., Banerji, N., & Nation, K. (2023). Effects of Target Age and Genre on Morphological Complexity in Children's Reading Material, Scientific Studies of Reading, DOI: 10.1080/10888438.2023.2206574

Hsiao, Y., Dawson, N.J., Banerji, N., & Nation, K. (2023). The nature and frequency of relative clauses in the language children hear and the language children read: A developmental cross-corpus analysis of English complex grammar. *Journal of Child Language, 50*(3), 555-580. doi:10.1017/S0305000921000957

Nation, K., Dawson, N. J., & Hsiao, Y. (2022). Book language and its implications for children's language, literacy, and development. *Current Directions in Psychological Science*, *31*(4), 375–380. doi: 10.1177/09637214221103264

**Dong, Y., Hsiao, Y., Dawson, N., Banerji, N., & Nation, K. The emotional content of children's writing: a data-driven approach. *Cognitive Science*, https://doi.org/10.1111/cogs.13423

**Hsiao, Y., Dawson, N., Banerji, N., & Nation, K. A corpus based developmental investigation of lexical richness and syntactic complexity in children's written stories. *Applied Corpus Linguistics*, https://doi.org/10.1016/j.acorp.2024.100084

SELECTED TALKS AND PRESENTATIONS (ACADEMIC AUDIENCE)

- Dawson, N.J. The power of words in children's stories: does lexical richness enhance children's narrative retelling and word knowledge? Society for the Scientific Studies of Reading, Queensland, Australia, July 2023.
- Nation, K. Emotion words in children's reading and writing: book language and its implications for social-emotional health. Society for the Scientific Studies of Reading, Queensland, Australia, July 2023.
- Nation, K. Book language and its implications for children's social-emotional learning. Workshop on language, literacy and mental health, University College London, June 2023.
- Nation, K. Mid-career Award Lecture: Becoming a Reader. Experimental Psychology Society conference, March 2022.
- Dawson, N.J. et al. Lexical richness and children's book language: Corpus explorations and a registered report. ReadingFest, University of Oxford (with national and international participants), March 2022.
- Zhang, M. et al. Effects of target age and genre on morphological complexity in children's reading materials: Establishing developmental statistics. ReadingFest, University of Oxford (with national and international participants), March 2022.
- Dawson, N.J. Characterising children's book language: Links to language and literacy. Royal Holloway, University of London (Language and Reading Acquisition group), March 2022.
- Hsiao, Y. et al. Lexical and syntactic features of children's book language. Online spoken presentation, Macquarie University Centre for Reading, Sydney, Australia, May 2021.

- Nation, K. The nature and content of children's book language: implications for language and literacy development. NAPLIC annual conference 2021: Language: The Bridge Across the Gap, May, 2021.
- Dawson, N.J. Exploring lexical and syntactic features of children's book language. Forum for Research in Literacy and Language, December 2020.
- Dawson, N.J. Characterising lexical richness in the language of children's books. LiFT group, Department of Education, University of Oxford, December 2020.
- Dawson, N.J. Characterising lexical richness in the language of children's books. Many Paths to Language (MPaL) conference, Max Planck Institute, Netherlands, October 2020.

SELECTED TALKS AND PRESENTATIONS (PROFESSIONAL AUDIENCE)

- Nation, K. Book language and its implications for children's language, literacy, and development. SPELD-New South Wales, Sydney, Australia, August 2023.
- Nation, K. Book language and its implications for children's language, literacy, and development. Dyslexia-SPELD Foundation, Perth, Australia, July 2023.
- Dawson, N.J. The Power of Books: What children learn about language through listening to stories. Yealmpstone Farm Primary School (INSET day talk to school and nursery practitioners/speech and language therapists), January 2022
- Nation, K. The nature and content of children's book language: implications for language and literacy development. NAPLIC annual conference 2021: Language: The Bridge Across the Gap, May 2021
- Dawson, N.J. Language and learning to read. Online as part of "psychology week" outreach event, November 2021

PUBLIC AND PROFESSIONAL ENGAGEMENT

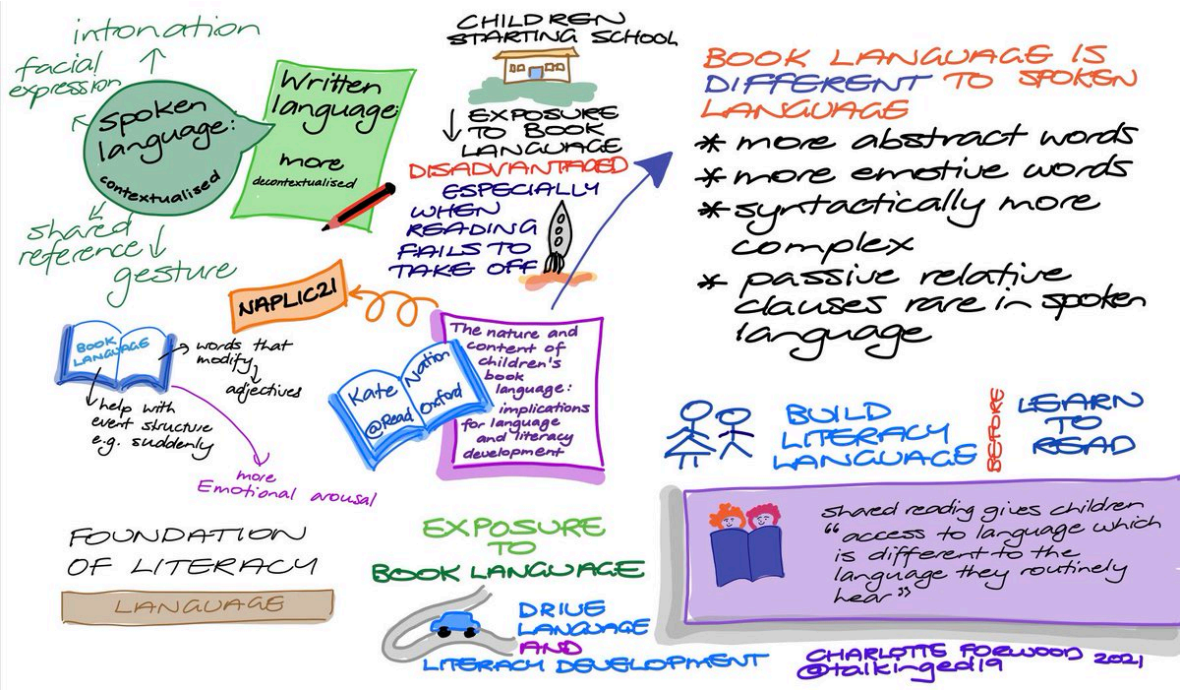
- 'Supporting Children with DLD to Access the English Curriculum'. Training webinar for teachers and teaching assistants, Moor House Research & Training Institute (delivered by MHRTI staff)

- A Book at Bedtime, Video for Oxford Sparks (providing scientific engagement and discovery for young people). Available at https://www.oxfordsparks.ox.ac.uk/videos/a-book-at-bedtime/
- Inspire Programme summer school academic taster lecture for Year 12 students, 'The Psychology of Language', St John's College, University of Oxford
- After dinner talk for secondary school teachers' annual residential course, 'Reading in adolescence: Research and implications', St John's College, University of Oxford

MEDIA AND SOCIAL MEDIA COVERAGE (BY OTHERS)

- Dixon, M. (2021, April 16). Pupils' post-lockdown language need a boost? Try a book. *TES Magazine*. https://www.tes.com/magazine/article/pupils-post-lockdown-language-need-boost-try-book
- Kinnane, D. Tips and resources to help school-aged children learn book language. Blog, links, and resources. https://www.banterspeech.com.au/free-tips-and-resources-to-help-your-preschooler-or-school-aged-child-learn-book-language-for-later-school-and-life-success/

Infographic produced by Dr Charlotte Forwood (while listening to Kate Nation's presentation on children's book language at the NAPLIC conference, 2021).



ACKNOWLEDGEMENTS

Thank you to the Nuffield Foundation who funded this work. The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds research that informs social policy, primarily in Education, Welfare, and Justice. It also funds student programmes that provide opportunities for young people to develop skills in quantitative and scientific methods. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute and the Nuffield Family Justice Observatory. The Foundation has funded this project, but the views expressed are those of the authors and not necessarily the Foundation. Visit www.nuffieldfoundation.org.

A special thank you to Eleanor Ireland, Programme Head for Education at the Nuffield Foundation, who has steered us with care and flexibility throughout, especially when the pandemic made our work nearly impossible. We are grateful for her on-going patience and enthusiasm, and for embracing the Registered Report format from the outset.

Thank you too to our Advisory Board members – the pandemic meant we couldn't meet as originally envisaged, but you have always been on hand to advise and guide. We thank Professor Charles Hulme and all at OxEd and Assessment for supporting our use of the Language Screen.

Enormous thanks to the many schools and nurseries that have supported this work. In amongst all else on your plates, you found time and space to host our project. We are never not inspired when we step inside your classrooms.

Thank you, of course, to the children and families who participated in our study. We've loved reading and hearing your stories, and we have learned a lot too. And our own stories and stimulus materials were so much the better for Neil Usher's illustrations, carefully and thoughtfully crafted to address our research questions; thank you Neil.

Our friends and colleagues at OUP made this work not just possible, but also enjoyable. A big thank you to Vineeta Gupta for the early days, and to the entire team since then: in particular, Nilanjana Banerji, Helen Freeman, Samantha Armstrong, and Rebecca Lawrence, and to everyone for sharing a little 500 Words magic with us.

As ever, research is a team effort. We are grateful to the team of students who have helped with data collection and transcription or built on some of our findings in their own undergraduate degree projects. We hope this taste of research inspires you for the future.

Finally, we thank out academic homes (the Department of Experimental Psychology and St John's College) for providing the physical, financial, and intellectual resources needed to conduct this research.

REFERENCES

Adams, C., Cooke, R., Crutchley, A., Hesketh, A., & Reeves, D. (2001). *Assessment of Comprehension and Expression 6-11*. GL Assessment.

Aslin, R. N., & Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning*, *64*(Suppl.2), 86–105. https://doi.org/10.1111/lang.12074

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Bus, A. G., Van IJzendoorn, M. H., & Pellegrini, A. D. (1995). Joint Book Reading Makes for Success in Learning to Read: A Meta-Analysis on Intergenerational Transmission of Literacy. *Review of Educational Research*, *65*(1), 1–21. https://doi.org/10.3102/00346543065001001

Cameron-Faulkner, T., & Noble, C. (2013). A comparison of book text and Child Directed Speech. *First Language*, *33*(3), 268–279. https://doi.org/10.1177/0142723713487613

Cassani, G., Grimm, R., Daelemans, W., & Gillis, S. (2018). Lexical category acquisition is facilitated by uncertainty in distributional co-occurrences. *PLoS ONE*, *13*(12), 1–36. https://doi.org/10.1371/journal.pone.0209449

Catts, H. W. (2018). The Simple View of Reading: Advancements and False Impressions. *Remedial and Special Education*, *39*(5), 317–323. https://doi.org/10.1177/0741932518767563

Dawson, N., Hsiao, Y., Tan, A. W. M., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*. https://doi.org/10.34842/5we1-yk94

Dawson, N., Hsiao, Y., Tan, A. W. M., Banerji, N., & Nation, K. (2023). Effects of Target Age and Genre on Morphological Complexity in Children's Reading Material. *Scientific Studies of Reading*, 1–28. https://doi.org/10.1080/10888438.2023.2206574

Hadley, P. A., Rispoli, M., Holt, J. K., Papastratakos, T., Hsu, N., Kubalanza, M., & McKenna, M. M. (2017). Input Subject Diversity Enhances Early Grammatical Growth: Evidence from a Parent-Implemented Intervention. *Language Learning and Development*, *13*(1), 54–79. https://doi.org/10.1080/15475441.2016.1193020

Hadley, P. A., Rispoli, M., & Hsu, N. (2016). Toddlers' Verb Lexicon Diversity and Grammatical Outcomes. *Language, Speech, and Hearing Services in Schools*, *47*, 44–58. https://doi.org/10.1044/2015

Hamilton, L. G., Hayiou-Thomas, M. E., Hulme, C., & Snowling, M. J. (2016). The home literacy environment as a predictor of the early literacy development of children at family-risk of dyslexia. *Scientific Studies of Reading*, *20*(5), 401–419. https://doi.org/10.1080/10888438.2016.1213266

Hills, T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory & Language*, *63*(3), 259–273. https://doi.org/10.1038/jid.2014.371

Hoff, E., & Naigles, L. (2002). How Children Use Input to Acquire a Lexicon. *Child Development, 73*(2), 418–433.

Hsiao, Y., Dawson, N. J., Banerji, N., & Nation, K. (2022). The nature and frequency of relative clauses in the language children hear and the language children read: A developmental cross-corpus analysis of English complex grammar. *Journal of Child Language*, 1–26. https://doi.org/10.1017/S0305000921000957

Hsu, N., Hadley, P. A., & Rispoli, M. (2017). Diversity matters: Parent input predicts toddler verb production. *Journal of Child Language*, *44*(1), 63–86. https://doi.org/10.1017/S0305000915000690

Lingwood, J., Billington, J., & Rowland, C. (2020). Evaluating the Effectiveness of a 'Real-World' Shared Reading Intervention for Preschool Children and Their Families: A Randomised Controlled Trial. *Journal of Research in Reading*, *43*(3), 249–271. https://doi.org/10.1111/1467-9817.12301

Mol, S. E., Bus, A. G., De Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent-child book readings: A meta-analysis. *Early Education and Development*, *19*(1), 7–26. https://doi.org/10.1080/10409280701838603

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. https://doi.org/10.1177/0956797615594361

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447–468. https://doi.org/10.1037/xge0000054

Nation, K. (2019). Children's reading difficulties, language, and reflections on the simple view of reading. *Australian Journal of Learning Difficulties*, *24*(1), 47–73. https://doi.org/10.1080/19404158.2019.1609272

Nation, K., Dawson, N., & Hsiao, Y. (2022). 'Book language' and its implications for children's language, literacy, and development. *Current Directions in Psychological Science*. https://doi.org/10.1177/09637214221103264

Noble, C., Sala, G., Peter, M., Lingwood, J., Rowland, C., Gobet, F., & Pine, J. (2019). The impact of shared book reading on children's language skills: A meta-analysis. *Educational Research Review*, *28*(October 2018), 1–32. https://doi.org/10.1016/j.edurev.2019.100290

Rastle, K. (2019). The place of morphology in learning to read in English. *Cortex*, *116*, 45–54. https://doi.org/10.1016/j.cortex.2018.02.008

Rastle, K. (2022). Word Recognition III: Morphological Processing. In *The Science of Reading: A Handbook* (2nd ed., pp. 102–120). Wiley Blackwell.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379. https://doi.org/10.1016/j.jml.2007.03.002

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech vocabulary development. *Child Development*, *83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, *79*, 1–39. https://doi.org/10.1016/j.cogpsych.2015.03.003

West, G., Snowling, M. J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H., & Hulme, C. (2021). Early language screening and intervention can be delivered successfully at scale: Evidence from a cluster randomized controlled trial. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. https://doi.org/10.1111/jcpp.13415