

**The power of the page: Comparing richness in text and talk  
during book sharing with two-year-old children**

Alvin W. M. Tan<sup>1</sup>, Kirsten Read<sup>2</sup>, Sophia Gamboa<sup>3</sup>, Janet Y. Bang<sup>4</sup>, and Virginia A. Marchman<sup>5</sup>

<sup>1</sup>Department of Psychology, Stanford University

<sup>2</sup>Department of Psychology, Santa Clara University

<sup>3</sup>Program in Human Biology, Stanford University

<sup>4</sup>Child and Adolescent Development Department, San José State University

<sup>5</sup>Department of Psychology, Stanford University

### Author Note

Alvin W. M. Tan  <https://orcid.org/0000-0001-5551-7507>

Kirsten Read  <https://orcid.org/0000-0002-1334-9610>

Janet Y. Bang  <https://orcid.org/0000-0002-6014-3009>

Virginia A. Marchman  <https://orcid.org/0000-0001-7183-6743>

Code and pre-processed data for this manuscript are available at [osf.io/q26jx](https://osf.io/q26jx). Raw book transcripts cannot be shared due to copyright restrictions. Raw speech transcripts are forthcoming on the CHILDES data repository.

Ethics approval was obtained from the Institutional Review Board of Stanford University (IRB #11109). All participants gave informed written consent before taking part in the study.

The authors declare no conflicts of interest.

We are grateful to the children and parents who participated and the research assistants who helped code and transcribe: Arlyn Mora, Mónica Munévar, Jessica Magallón, Nadia Segura, Shriya Anand, Marisol Rodriguez, Maria Lopez, Stephen Lopez, Jesús Esquivel-Barrientos, Laura Jonsson, Kalpana Gopalkrishnan, Maribel Mercado, Tami Alade, Jaqueline De Paz-Romero, Lesly Leon, Alice Articia, Julia Briones-Avila, and Elizabeth Sanchez. We also thank our colleagues for helpful discussions including Sarah Surrain, Michael C. Frank, Talya Gómez, and the staff of the Language Learning Laboratory at Stanford University.

This work was supported by grants from the National Institutes of Health (R01 DC008838, R01 HD092343, 2R01 HD069150), the Schusterman Foundation, the David and Lucile Packard Foundation, the Bezos Family Foundation, and the Stanford Maternal and Child Health Research Institute.

The authors made the following contributions. Alvin W. M. Tan: Conceptualization, Methodology, Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing; Kirsten Read: Conceptualization, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; Sophia Gamboa: Conceptualization, Methodology, Investigation;

Janet Y. Bang: Conceptualization, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; Virginia A. Marchman: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing, Supervision

Correspondence concerning this article should be addressed to Alvin W. M. Tan, Department of Psychology, Stanford University, 450 Jane Stanford Way Building 420, Stanford, California 94305, USA, Email: [tanawm@stanford.edu](mailto:tanawm@stanford.edu)

### Abstract

Sharing books with young children has long been associated with positive outcomes for children's language and literacy outcomes. Studies suggest that caregiver language during these book-sharing interactions is linguistically richer than speech directed to young children (target-child-directed speech; tCDS) during other everyday activities. However, few studies have directly compared the linguistic richness of speech in book sharing and other activities within the same families. In this study, tCDS was sampled from day-long naturalistic recordings to investigate the linguistic richness of speech directed to 24-month-old children from English- and Spanish-speaking families in the U.S. We compared four sources of speech: book text, read-aloud speech during book-sharing interactions, spontaneous speech during book-sharing interactions, and spontaneous speech during other everyday activities. We first used a mega-transcript approach combining across families ([Dawson et al., 2021](#); [Montag et al., 2015](#)) and compared the four sources of language on type-token ratio curves. We next conducted a more fine-grained analysis to look within families, comparing the four sources of speech in terms of lexical (lexical diversity, lexical density, lexical frequency, and contextual diversity) and grammatical (mean length of utterance in words and the proportion of complex utterances) measures associated with linguistic richness. Across both approaches and in both language groups, speech based on text was lexically- and grammatically-richer than both sources of spontaneous speech. These results indicate that the power of the proposed linguistic richness of speech during book-sharing interactions stems from features of the text on the page.

*Keywords:* book sharing, linguistic richness, extratextual talk, child-directed speech

## **The power of the page: Comparing richness in text and talk during book sharing with two-year-old children**

### **Introduction**

Reading aloud with young children is largely considered to be beneficial for children's language development. Decades of research has demonstrated that the frequency and quality of shared book reading are positively correlated with short- and long-term increases in receptive and expressive language skills (e.g., [Arterberry et al., 2007](#); [Debaryshe, 1993](#); [Farrant & Zubrick, 2012](#); [Flack et al., 2018](#); [Karrass & Braungart-Rieker, 2005](#); [Raikes et al., 2006](#); [Sénéchal et al., 1996](#); [Sénéchal et al., 2008](#); [Sénéchal & LeFevre, 2014](#)). General benefits of frequent shared book reading may appear obvious because book-sharing episodes are concentrated periods of language use during which caregivers read book texts aloud. However, caregivers are also likely to talk about what they are reading—for example, this ‘extratextual’ talk may include comments on the characters, questions, or language used to manage the activity. Few studies have directly compared the richness of the various types of talk that occurs during book-sharing episodes. In this study, we explore speech directed to 2-year-old children during naturally occurring book-sharing episodes in two diverse samples of English- and Spanish-speaking families in the U.S. Our goal is to explore various features of book-text talk and spontaneous talk to understand the features of the types of language used in these interactions.

### **Book sharing as a source of “rich” language**

While many everyday activities include potentially linguistically rich adult-child conversations, it is generally claimed that shared book reading exposes children to “richer” language relative to other adult and child interactive contexts. Indeed, quasi-experimental studies of the speech heard by young children have found that English-speaking caregivers in the U.S. use more words in total ([Gilkerson et al., 2017](#); [Hoff-Ginsberg, 1991](#)), words per minute ([Soderstrom & Wittebolle, 2013](#); [Tamis-LeMonda et al., 2019](#)), lexically diverse words ([Hoff-Ginsberg, 1991](#); [Tamis-LeMonda et al., 2019](#)) and complex sentences ([Hoff-Ginsberg, 1991](#)) during shared book reading activities than during other contexts (e.g., mealtime, toy play). These findings have led to

an overarching claim that the language used during shared book reading with young children is especially rich.

The “richness” of child-directed speech (CDS) can be measured in many ways. First, there may be important differences across contexts in speech quantity (e.g., total number of words heard, or rate of speech). Quantity may be important because simply hearing more words, or more words per minute, can enrich the environment for a child learning words (e.g., [Hoff-Ginsberg, 1991](#)). Second, the content of child directed speech (e.g., level of vocabulary, abstractness of ideas) has also been considered. For example, rich language is speech that is high in lexical density (i.e., content words), uses low frequency (e.g., “rare” or “sophisticated”) words, and contains complex sentences (e.g., [Montag, 2019](#); [Weizman & Snow, 2001](#)). If the language that children are exposed to during shared book reading is measurably rich in these many ways, then it follows that reading aloud with a young child can be a useful way for caregivers to provide high-quality speech to young children.

However, it remains unclear precisely why the language during book-sharing is so rich. Unlike other activities, book sharing is unique in that what caregivers say is a reflection of both the words on the page (i.e., the text itself) and the “extra” talk, i.e., the talk about the content of the book (e.g., pictures, new words, letters) or even the activity of book sharing itself (e.g., listening, turning the page). Much of the literature has focused on the language that a child could hear—the text printed within the books themselves if it were read aloud verbatim. Notably, the text of children’s books in English has been found to be “richer” than other sources of speech directed towards children. Montag et al. ([2015](#)) compared the text of 100 popular American children’s books in English (63,103 words) to 4,432 adult-child conversations (6.5 million words) from American English corpora of transcribed speech directed at children under 5 years old. More recently, Dawson et al. ([2021](#)) compared the text of 160 popular British children’s books in English to 1,616 samples of child-directed speech from UK corpora transcribed from everyday interactions with children under 6 years old. Both studies found greater lexical richness in children’s book texts compared to corpora of child-directed speech. Measures of lexical richness

included lexical diversity (Dawson et al., 2021; i.e., type/token ratios at all sample sizes; Montag et al., 2015), lexical density (i.e., the proportion of content words out of total words; Dawson et al., 2021), and lexical sophistication (i.e., the rate of occurrence of rare words; Dawson et al., 2021; Massaro, 2015). Studies have also found that children's book texts tend to be more grammatically complex compared to transcribed CDS from a variety of child-centered activities, measured by the frequency of passive sentences and relative clauses (Montag, 2019), as well as canonical (i.e., comprising subject-verb-[object]) and complex (containing two or more lexical verbs) sentence constructions (Cameron-Faulkner & Noble, 2013).

These corpora-based studies highlight that the text on the pages of popular English language children's books is, by multiple measures, "richer" than the speech captured in recorded adult-child conversations in everyday contexts. When parents are given books by researchers to read aloud with their young children, they tend to read all the text verbatim (e.g., Read et al., 2021; Stoops & Montag, 2024). However, caregivers engaged in more natural shared book reading with a young child at home may only read a subset of the words and sentences printed in the books they share, and they may at times simplify the vocabulary, or grammar, as well as paraphrase, replace, expand on or even translate the texts that are read aloud. Thus, in order to understand the richness of the language actually heard by a child during shared book reading, we need to measure not just the text in the books, but the texts of the books that actually get read aloud during everyday, natural settings (e.g., Ece Demir-Lira et al., 2019; Stoops & Montag, 2024).

A second important source of talk during book sharing is the spontaneous commentary and conversation that occurs alongside the reading of the text itself. Not only do children hear the text of the book read aloud, readers also tend to elaborate and engage children in spontaneous extratextual commentary and conversation prompted by the book and the shared reading activity (Fletcher & Reese, 2005; Hindman et al., 2014; Mol et al., 2008; Read et al., 2023). For example, caregivers may talk about the story or book itself (e.g., "Do you see the cat"? "This is the cover."). Some talk can help children elaborate or connect the book to their own experiences (e.g., "Oh, you've tried zucchini before too, remember?"), and some talk, like in any other activity, may be

unrelated (e.g., "Is that the doorbell?"). In their 38-study meta-analysis, Flack et al. (2018) found that the use of dialogic reading styles, defined as 'adding something to a verbatim text reading,' such as extratextual repetitions, explanations or definitions of words, results in children learning 1.22 more words than non-dialogic reading (i.e., reading the text aloud verbatim). While there is a good deal of variability in the amount, the quality, and even the relevance of this extratextual spontaneous talk during shared book reading (e.g., Price et al., 2009; Read et al., 2023), it is certainly a common part of book sharing in children's everyday environments. Thus, it is important to include this spontaneous extratextual talk alongside the text that is read aloud in our comparison of shared book reading language to other activities.

At the same time, there is mixed evidence regarding whether spontaneous talk during book sharing is different from caregiver speech to children during other activities. Recent studies with English-speaking families have suggested that spontaneous speech during shared book reading, like the book texts themselves, provides forms of richer language compared to other activities. With a sample of 36 English-speaking families in the U.S. with 3.5-year-olds, Gilkerson et al. (2017) found that parents and children spoke more words and used more conversational turns per minute during shared reading time than during other non-reading times. Similarly, Soderstrom and Wittebolle (2013) found that, in Canada, both adult and child rate of talk at home and at daycare was highest during shared reading time compared to other structured and unstructured activities in 1- to 2-year-old toddlers. Moreover, Tamis-LeMonda et al. (2019) found that U.S. mothers used more words overall and used more lexically-diverse words during book sharing compared to talk during other kinds of activities, such as object play or feeding, with their 13-month-olds in their homes. In contrast, Weizman and Snow (2001) found that sophisticated words were used with greater probability during mealtime and playtime than during book reading. Similarly, Bang et al. (2022) found that book sharing was likely to include more words/minute in English, with similar trends in Spanish, but that the increased rate of speech did not necessarily reflect higher levels of lexical diversity in English or Spanish (i.e., types per minute) than talk in other child-centered contexts. Thus, different methods across studies have led to different



conclusions about whether the talk outside of the text read aloud during shared reading really is decidedly more rich than similar kinds of CDS during other activities.

Beyond the amount of talk or measures of lexical diversity and richness, other studies have focused on measures of sentence-level richness and grammatical complexity of the book reading talk compared to adult-child talk in other child-centered contexts, again, with mixed results. Ece Demir-Lira et al. (2019) found that, within a sample of 47 parent-child natural interactions recorded at 4 age points between 14 and 30 months, talk during shared book reading was both more lexically diverse and more grammatically complex than talk in other settings, as measured by mean length of utterance (MLU) and number of unique verbs per utterance. In another within-dyad comparison, Noble et al. (2018) found more complex language was produced by 43 parent-child dyads during extratextual talk during book sharing compared to toy play, as measured by proportion of Subject-Verb (SV) sentences and complex constructions (i.e., sentences with two or more lexical verbs). However, they also found adults used more questions, another complex sentence form, during toy play compared to book sharing.

Despite some mixed findings, prior studies suggest an overall richer language environment, with opportunities to hear more words, more word types, more sophisticated words, and more grammatically complex sentences, during shared book reading compared to during other types of child-centered activities. While many studies compared either book text or the spontaneous speech that occurs during shared reading with child directed speech during other activities, it is important to note that shared reading is an opportunity to hear (and learn from) both the texts that are read aloud and the conversation around those texts. With the exception of Ece Demir-Lira et al. (2019) and Stoops and Montag (2024), little previous research has separated out these different sources of language within the shared reading context. Understanding better where the rich language during book sharing originates—just within the book text itself, or also within the extratextual verbal interactions that a book text may instigate—can help us more clearly understand the potential benefits of the whole book sharing experience for a young child.

### **Broadening our view of caregiver talk during book sharing**

In exploring these issues, we seek to go beyond earlier work in two key ways. First, we include book-sharing interactions extracted from day-long recordings, where we have the opportunity to capture caregiver-child conversations as they naturally occur during everyday activities with books that families actually have on hand ([Gilkerson et al., 2017](#); [Soderstrom & Wittebolle, 2013](#)). Prior data on caregiver and child book-sharing interactions typically derive from short audio- or video-recordings in the home or laboratory that are often planned events with a particular book. While informative, these types of activities do not capture the natural individual family variance around what books are read (or whether books are read) throughout a typical day in the home. By examining everyday book-sharing experiences using day-long recordings, we can obtain a more representative picture of variation in the ways that caregivers engage verbally with their young children during book sharing ([N. J. Anderson et al., 2021](#)).

Second, the present study aims to capture the experiences of children from a broader, more representative American sample by including both primarily English- and Spanish-speaking families in the Western U.S., who spanned a range of demographic, sociolinguistic, and sociocultural backgrounds. Prior studies have typically included families from higher-educated backgrounds and who were English-speaking in the U.S. This subset of families in the U.S. limits generalizability to families from diverse socioeconomic and sociocultural backgrounds, yielding a constrained, incomplete, and possibly biased sample ([Blasi et al., 2022](#); [Singh & Rajendra, 2024](#)). English-speaking families in the current sample were recruited several hundred miles from research centers where we could identify more families who lived in primarily monolingual English-speaking rural communities than is typical of higher-socioeconomic university-based samples (REF-masked). We also included Spanish-speaking families who were recruited as part of a community outreach program. Spanish-speakers in the U.S. represent a heterogeneous population. Our sample includes many families who were recent immigrants to the U.S. from countries where Spanish was the dominant language—most typically, Mexico. Families primarily spoke Spanish in the home. Thus, the sample of book texts that they have on hand and choose to

read aloud with their children differs from those used in previous quantitative analyses of the language available within children's books. Most notably, the sample of books for Spanish-speaking families contained texts written in Spanish, as well as English, including some bilingual English–Spanish books. By including a range of language backgrounds, this study presents a more inclusive and representative sample of modern Western American families reading with young children. We examine general trends across both samples, but also explore language group as a potential moderator.

### **The present study**

We analyzed samples of naturally occurring activities extracted from day-long audio recordings in English- and Spanish-speaking families with two-year-old children. In each family, we identified the six densest 10-minute segments of target child-directed speech (tCDS) and identified all activities in which the talk occurred (e.g., book sharing, play, mealtimes). Our approach was as follows.

First, we compared features of the book text and actual speech to young children using multiple sources: (1) full texts of the books that were read aloud, (2) speech from those book texts that was read aloud by caregivers, (3) spontaneous, extratextual tCDS that occurred during book-sharing sessions, and (4) tCDS during all other non-book-sharing activities. Second, we used two different analytic approaches to the data. In the first set of analyses, we adopted an approach from Montag et al. (2015) and Dawson et al. (2021), in which transcripts from each source are concatenated across families to form a “mega-transcript” for each source. Our goal was to explore whether our smaller, more ecologically-valid and varied sample of book texts would also show that book texts were more lexically rich than spontaneous speech of various sorts (e.g., Dawson et al., 2021; Montag et al., 2015). Note that this mega-transcript approach is unable to account for by-family and by-transcript variability, for example, some caregivers may produce more unique types on average than others.

Therefore, a second set of analyses applied a more fine-grained approach in which we used mixed effects regressions to model differences in each of six richness measures across language

sources, while incorporating family- and transcript-level random effects. We compared the “richness” of speech across language sources using a comprehensive survey of multiple measures informed by prior research, capturing lexical richness (e.g., lexical diversity, semantic density, and vocabulary sophistication) and grammatical complexity. This approach aligns more with previous studies which have explored multiple measures across different activity types (e.g., [Hoff-Ginsberg, 1991](#)). By classifying caregiver speech into read-aloud text, spontaneous tCDS during book-sharing, and spontaneous tCDS during other activities, we could distinguish between a few different scenarios regarding the lexical and grammatical richness of tCDS. In the first scenario, the richness of book text also entrains rich extratextual talk, and therefore spontaneous tCDS during book-sharing would be richer than that during other activities. In the second scenario, the richness of spontaneous speech is not particularly influenced by the activity, and thus spontaneous tCDS during book-sharing may be more similar to other tCDS. In the third scenario, caregivers have relatively low fidelity when reading books, such that even the read-aloud text is less rich than the book text itself. The regression approach would help to determine which of these scenarios best characterizes caregiver speech in book-sharing and other activities.

## Method

### Participants

Participants (English-speaking:  $n = 22$ ; Spanish-speaking:  $n = 20$ ) were part of a larger study in which day-long naturalistic recordings were collected from English- and Spanish-speaking families living in the U.S. (REF-masked). In that larger study, participants (English:  $n = 45$ ; Spanish:  $n = 45$ ) were recruited between 2010 and 2013 through birth records in several sites in Northern California; recruitment information is published elsewhere (REF-masked). Eligible participants were asked to conduct a day-long naturalistic home recording using the LENA (Language Environment Analysis) system ([Gilkerson & Richards, 2008](#)) when their child was 2 years old. For each family, we then extracted and coded the six 10-minute segments (~1 hour) of their recording that were characterized by the densest tCDS. Because book-sharing was our focus, only those English- and Spanish-speaking families who had

at least one episode of book sharing during their densest hour of tCDS were included in the current analyses, representing about half of the larger sample in each group.

Table 1 presents the demographic characteristics of the samples. All children were approximately 2 years old. In both samples, there were relatively more girls than boys. On average, the mothers in the English-speaking families had ~3 years of post-high school education, although there was considerable variation. On average, mothers in the Spanish-speaking sample received high-school education; again, there was considerable variation. We explored the two groups separately, focusing on general trends that held across groups, but also considered language as a potential moderator when the groups were later combined.

**Table 1**

*Sample demographics of the children and families from the English- ( $n = 22$ ) and Spanish- ( $n = 20$ ) speaking families who engaged in any book sharing during their day-long audio recording.*

	English		Spanish	
	<i>M</i> (SD)	Range	<i>M</i> (SD)	Range
Age (months)	24.0 (0.8)	23–25	25.5 (0.5)	24–26
% male	36.4		45.0	
Maternal education (years)	15.3 (1.7)	12–18	12.3 (3.5)	6–18
Birth order				
1	12		7	
2	6		6	
3	3		5	
4+	1		2	
Maternal time in the U.S. (years)			13.5 (7.0)	5–34

### Activity coding

After data collection, the densest six 10 minute segments of tCDS was extracted from each LENA recording for further analysis. To do so, consecutive 5-minute segments were concatenated into 10-minute audio segments which were then sorted in descending order based on the automated adult word counts (AWC) provided by the LENA software. Trained human listeners classified each 10-minute segment as either tCDS or non-tCDS. Segments were classified as tCDS if ~70% of speech was directed to the target child who was wearing the LENA recorder. Listeners classified segments until six tCDS segments were identified, resulting in ~1 hour of each family's densest hour of tCDS.

Trained coders then annotated each primarily-tCDS segment for all the activities in which caregivers and children were verbally engaged, noting the start and stop time of each activity. Activities were classified as either child-centered (book sharing, dressing/routines, play, mealtimes, unstructured conversation) or adult-centered (e.g., adult is talking with the child while they are preparing a meal). All activities were identified using the content of the caregiver speech (e.g., "Let's read a book!", "Do you want to eat your peas?"), prosodic features (e.g., book-sharing register), or other environmental cues (e.g., sound of pages turning, pans clanking). For the current analyses, the speech occurring during book-sharing episodes was analyzed separately from that occurring during all other primarily-tCDS child-centered and adult-centered activities. Additional details regarding recruitment and activity coding can be seen elsewhere (REF-masked).

To ensure our sampling procedure was effective at capturing all book-sharing episodes, we performed randomized listening checks. In the English-speaking sample, we listened to 20% of the audio recordings that originally did not have any book-sharing episodes in the densest hour of speech. Only one additional book-sharing episode was identified, representing about 1% of all speech, and was not further analyzed. In the Spanish-speaking sample, we listened to 20% of the recordings from 20% of non-book-sharing families (i.e., 5 families). No instances of book sharing were identified. This evidence suggests that our original method was effective at identifying the vast majority of the book-sharing episodes existing in our speech samples.

For the current analyses, all book-sharing episodes were further curated to ensure that each episode was captured in its entirety. Notably, if any book-sharing episode extended beyond the original 10-minute segment (i.e., either the start or the end of a book-sharing episode did not fall within the segment), a trained research assistant listened to identify the true start and/or end time of that episode. These additional portions of the recording were then appended to the original segments.

Table 2 shows that each family contributed, on average, 1 to 2 book-sharing episodes, although some families had as many as 6 episodes. Families may have read more than 1 book per episode. The duration of each book-sharing episode was 5 to 6 minutes on average, but there was wide variability, with some episodes lasting only 1 minute to others lasting nearly 18 minutes.

**Table 2**

*Descriptive statistics of book-sharing episodes in English- ( $n = 22$ ) and Spanish- ( $n = 20$ ) speaking families.*

	English		Spanish	
	<i>M</i> (SD)	Range	<i>M</i> (SD)	Range
Number of book-sharing episodes	2.14 (0.94)	1–5	1.85 (1.27)	1–6
Duration of book-sharing episodes (min)	5.54 (2.73)	1.85–10.83	6.52 (3.46)	1.15–17.50

### Identifying book texts

Research assistants identified the books shared with the child using information spoken by the caregiver in the recording, including: 1) title of the book, 2) author name, 3) read aloud text that could be matched to the actual book, and/or 4) specific details of the pictures that the caregiver narrated. Across all English-speaking families ( $n = 22$ ), a total of 69 books were read during the episodes, of which 56 titles (81%) were confirmed for 21 families ( $M = 2.7$  titles per family, range = 1–8). Of the 21 families, there were 4 families where caregivers at some point read text where we could not identify the book title; these utterances were included as read-aloud text

in our analyses and were not represented in the book texts (66 utterances total). These read aloud utterances with unknown book titles represented a small percentage of the read aloud speech in our English sample (66 utterances / 2395 read aloud, 2.76%). Across all Spanish-speaking families ( $n = 20$ ), a total of 45 books were read, of which 30 titles were confirmed (67%) representing 15 families ( $M = 2.0$  titles per family, range = 1–6). Of the 15 families, there were 5 families where the caregivers at some point read text where we could not identify the book title; these utterances were included as read-aloud text in our analyses and were not represented in the book texts (108 utterances total). These read aloud utterances also represented a small percentage of read aloud speech in the Spanish sample (108 utterances / 1655 read aloud, 6.53%). All confirmed book titles were obtained online (e.g., YouTube read-alouds), at our local library, or purchased.

Books were then categorized in two ways. First, books were grouped according to the language of the text as English, Spanish or bilingual (i.e., both English and Spanish text). For the English-speaking families, the language of all books was English, whereas, the language of the books read by the Spanish-speaking families was Spanish (16/30, 53%), bilingual (11/30, 37%), or English (3/30, 10%). Second, books were categorized into type based on the structure of the text as non-prose or prose. Non-prose books were those in which the text consisted of naming routines (e.g., vocabulary or labeling books), repeated sentence frames (e.g., “Good night X”), or rhyme schemes (e.g. nursery rhymes). Prose books were those in which the text consisted of complete sentences that did not meet the naming, repeated frame, or rhyming criteria. A complete list of identified books can be found on our OSF repository ([osf.io/q26jx](https://osf.io/q26jx)).



## Transcription

Figure 1 outlines the four sources of language analyzed here. First, all book texts were transcribed in their entirety from all confirmed books that were read (in whole or in part) during the episodes. Second, read aloud text consisted of all utterances where it could be determined that the caregiver was reading the book text during book-sharing episodes. Coders used a variety of cues to identify read-aloud speech, including environmental sounds (e.g., pages being turned), a caregiver mentioning the book or its title, and/or the speech occurring in a book-reading register or



**Figure 1**

*Examples of the four sources of language: (1) book text, (2) read aloud text:, (3) spontaneous speech during book sharing, and (4) spontaneous speech during other activities, for both the English and Spanish samples.*

BOOK TEXT	SPEECH DIRECTED TO CHILDREN	
1. <b>Book text</b>	2. <b>Read aloud (during book sharing)</b> Text read aloud from the book 3. <i>Spontaneous-book (during book sharing)</i> Spontaneous tCDS during book sharing	4. <i>Spontaneous-other (during other activities)</i> Spontaneous tCDS during other activities
English / Inglés		
<p><b>Being together is fireside snuggles.</b> It's magical stories you share. It's hearing the rain patter pat on the windows. Squish squashed in your favorite chair.</p> 	<p>MOT: <i>yeah it's raining.</i> MOT: <i>but not at our house, huh?</i> <b>MOT: it's hearing the rain.</b> CHI: <i>yeah.</i> MOT: <b>pitter pat on the windows.</b> CHI: <i>yeah?</i> <b>MOT: squish squashed in your favorite chair.</b> CHI: <i>yeah.</i> MOT: <i>is it raining in the book, in the story?</i> CHI: <i>no.</i> MOT: <i>it is in the story.</i> MOT: <i>but is it raining outside our house?</i> CHI: <i>no.</i> MOT: <i>no.</i> MOT: <i>it's kinda sunny outside our house.</i></p>	<p>MOT: <i>[child name], did you tell Papa what we did this morning?</i> CHI: <i>baby.</i> MOT: <i>we woke up and we went outside and lit a fire and we chanced [pet name] and [grandmother name] around, huh?</i> CHI: <i>&amp;=laughs.</i> GFA: <i>did you have a fire this morning?</i> CHI: <i>yeah?</i> MOT: <i>squish squashed in your favorite chair.</i> CHI: <i>ball [x 2]!</i> MOT: <i>and you threw the ball for [pet name].</i> MOT: <i>and you watched him do hot_laps@g.</i> CHI: <i>hop.</i> MOT: <i>yup.</i></p>
Spanish / Español		
<p><b>¡Ahora es la hora de divertirse!</b> Clifford flota en una barra de jabón.</p> 	<p><b>MOT: ahora es la hora de divertirse.</b> <b>MOT: Clifford flota en la barra de jabón.</b> MOT: <i>ya?</i> MOT: <i>quieres seguir jugando?</i> CHI: <i>xxx quipo [: Clifford].</i> MOT: <i>Clifford.</i> MOT: <i>qué hace Clifford?</i> MOT: <i>qué hace Clifford, eh?</i> CHI: <i>billano [: cepillando] lo(s) (d)ientes.</i> MOT: <i>cepillando los dientes?</i> ADF: <i>cuál?</i> CHI: <i>eh boca [x 2]!</i> MOT: <i>la boca también.</i> MOT: <i>&amp;=laughs.</i></p>	<p>FAT: <i>qué haces?</i> MOT: <i>ten pápi.</i> CHI: <i>no sé.</i> CHI: <i>xxx.</i> FAT: <i>métete [x 2].</i> FAT: <i>sí [x 2].</i> CHI: <i>&amp;=protests.</i> FAT: <i>mira [x 2].</i> FAT: <i>esa foto de quién es?</i> FAT: <i>esa es la foto de Tía?</i> FAT: <i>esa allí es la foto de Tía.</i> FAT: <i>esa allí es foto que tiene en el cuadro.</i> FAT: <i>de quién es?.</i> FAT: <i>yo creo que es Tía.</i></p>

*Note:* Book text and the caregiver's (MOT = mother; FAT = father) read aloud speech is bolded. The caregivers' spontaneous tCDS is italicized. Child speech (CHI) is in regular font. Book text in English is from *When We're Together* by Claire Freedman. Book text in Spanish is from *Clifford y la Hora del Baño* by Norman Bridwell.

tone. All read aloud text was verified using the available texts; minor modifications in wording were acceptable, e.g., omissions or substitutions of words. Next, spontaneous speech during book sharing consisted of all utterances that were spontaneously produced and directed to the child during the book-sharing episodes, but were not identified as book text. Finally, spontaneous-other speech was all tCDS produced during all non-book-sharing episodes (e.g., play, diaper changing). More information on transcription and coding can be found in our manual (REF-masked).

All transcriptions followed CHAT conventions ([MacWhinney, 2000](#)). Per these conventions, utterance breaks during verbal interactions (including read-aloud text) were determined using intonational features, such as pauses, conversational turns, exclamations, and questions. Utterance breaks in book text were determined by punctuation and/or line breaks as printed on the page.

### Analysis 1

Our first set of analyses used a mega-transcript approach, following Montag et al. ([2015](#)) and Dawson et al. ([2021](#)). These earlier studies demonstrated that book text has greater lexical diversity than tCDS when considering the number of unique types and type–token ratio curves across a range of token window sizes. Critically, these studies used curated book texts from lists of popular books and compared those texts to the caregiver speech that occurred in transcripts available from CHILDES. In contrast, our analyses pulled both sources of text from the same samples of families, thus ensuring that the speech differed only by the specific factor of interest, i.e., source. Additionally, our analyses parsed two separate sources of speech produced by caregivers, read aloud text and spontaneous-book speech, which had not been examined in previous work.

### Measures and analytic strategy

In each language, we concatenated the text from each source (book text, read-aloud text, spontaneous-book, spontaneous-other) to form four mega-transcripts that represented all of the language sources that occurred across all participants. For each mega-transcript, we then drew 100 random samples at sample sizes ranging, in 100 token increments, from 100 to 7,700 tokens

(this number was constrained by the size of the smallest source: Spanish spontaneous–book). At each sample size, we then averaged over the 100 samples to derive the mean type count, i.e., the mean number of different words, per sample size.

### **Results: Analysis 1**

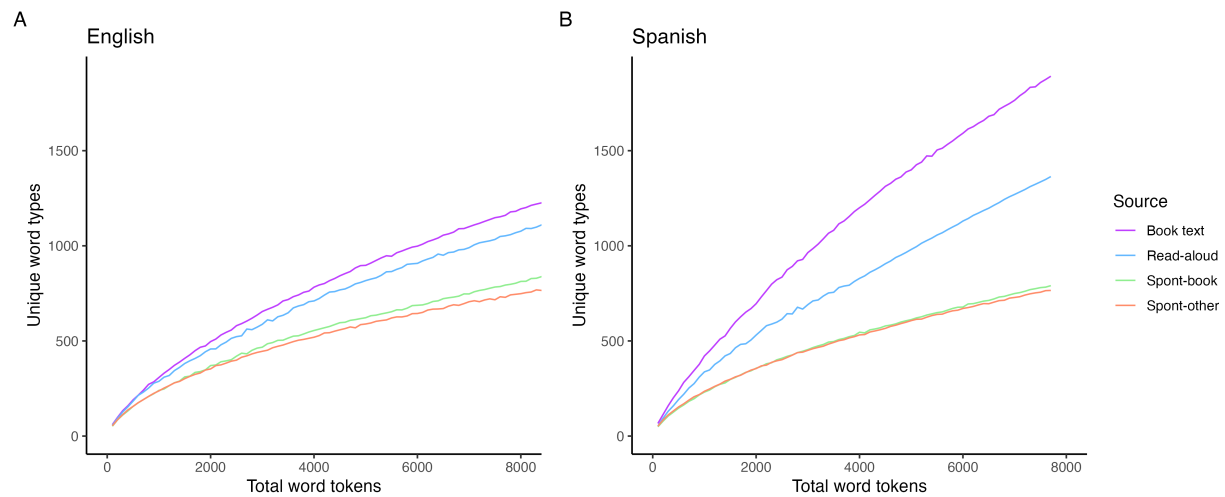
Figure 2 shows the number of unique word types as a function of sample size for each mega-transcript. Each curve represents the number of unique types by sample size for each of the four sources (book text, read aloud text, spontaneous-book, and spontaneous-other) in English- and Spanish-speaking families. Note that the differences among the source types become more pronounced as sample size increases, as found in earlier studies (e.g., Montag et al., 2015). Moreover, replicating Montag et al. (2015) and Dawson et al. (2021), we found that book text had more unique word types than speech from spontaneous–other at every sample size. Because we also explored text in read-aloud contexts, we were also able to demonstrate that book texts and read-aloud text tended to travel together, albeit read-aloud text had type counts that were somewhat lower across the range of sample sizes than those in book texts. We also demonstrated that these two text-based sources were more lexically diverse than spontaneous speech to children during both book and non-book activities. Note that while the magnitude of the differences across sources in the English- and Spanish-speaking samples varied, the relative orderings were similar.

### **Discussion: Analysis 1**

These results replicated the earlier finding that book texts tended to have higher lexical diversity, as measured in terms of both number of unique types and TTR, than tCDS produced in non-book-sharing contexts for any given token sample size (Dawson et al., 2021; Montag et al., 2015). This trend generally held across both English and Spanish families. Further, we demonstrated that the lexical diversity of read-aloud texts was similar to that of book texts and that the two sources of spontaneous speech (spontaneous-book and spontaneous-other) were generally lower in lexical diversity yet quite similar to each other. Similar, albeit not identical, patterns were found in the English- and the Spanish-speaking samples. To further investigate the differences across language sources in a broader range of lexical and grammatical measures, allowing a finer

**Figure 2**

*Number of unique word types as a function of token sample size per mega-transcript type in (A) English-speaking ( $n = 22$ ) and (B) Spanish-speaking ( $n = 20$ ) families.*



level of detail, we moved to a by-transcript approach, which also allowed us to control for inter-family variation.

### Analysis 2

In our second set of analyses, we used a by-transcript approach that allowed us to derive estimates of lexical and grammatical richness measures per episode, and then compare sources of language along these measures. This approach allowed us to test hypotheses statistically, to control for inter-family variation, and to adopt a more comprehensive approach across multiple dimensions of richness. For lexical richness, we measured lexical diversity, lexical density, and two indices of lexical sophistication (lexical frequency and contextual diversity); for grammatical richness, we measured the mean length of utterance in words (MLU-w) and the proportion of complex utterances.

For Spanish-speaking families, the large majority of speech across all activities was in Spanish (95%; see Bang et al., 2022). However, at times when English or bilingual books were read, they may have used both English and Spanish. All text/speech in each source, regardless of the language used, was combined into a single transcript for the computation of all measures.

## Measures and analytic strategy

*Lexical diversity.* Single type–token ratio measures are insufficient metrics of lexical diversity as they are sensitive to the length of the document (see e.g., [Hess et al., 1986](#)); thus, we calculated the measure of textual lexical diversity (MTLD) for each transcript. MTLD is defined as the mean length of sequential word strings in a text that maintain a type-token ratio that exceeds some predetermined threshold ([McCarthy, 2005](#); for further discussion on the length sensitivities of lexical diversity measures, see also [Fergadiotis et al., 2015](#); [McCarthy & Jarvis, 2010](#); [Stills, 2016](#)).<sup>1</sup> This analysis was conducted using the `koRpus` package in R ([Michalke, 2021](#)), with a type-token ratio threshold of .72. A higher score on MTLD indicates that a transcript contains more sequences of text that have a high number of different words.

*Lexical density.* Following Dawson et al. ([2021](#)), lexical density refers to the proportion of all tokens in the transcript that were coded as lexical, rather than non-lexical. Lexical tokens were defined as nouns (excluding proper nouns and pronouns), adjectives, verbs (excluding modal and auxiliary verbs), and deadjectival adverbs. All other tokens were coded as non-lexical. We calculated lexical density by dividing the number of lexical tokens by the number of total tokens in each transcript ([Berman & Nir, 2010](#); [Strömquist et al., 2002](#)). A higher score on lexical density indicates that a transcript contains relatively more contentful words.

*Lexical frequency.* We measure lexical sophistication in two ways. First, following Dawson et al. ([2021](#)), we calculated the log frequencies of each word token produced by caregivers in our transcripts that also occurred in a reference subtitle corpus: SUBTLEX-US for English ([Brysbaert & New, 2009](#)) and EsPal for Spanish ([Duchon et al., 2013](#)), operationalized as the number of occurrences of each token per million words in the reference corpus. These subtitle corpora had a coverage of 87.7% of all tokens for English and 81.9% of all tokens for Spanish. We then calculated the mean of the log reference frequencies for all tokens in each transcript,

---

<sup>1</sup> Note that MTLD is generally not recommended for texts of < 100 tokens ([Koizumi, 2012](#); [McCarthy & Jarvis, 2010](#)); however, removing transcripts of < 100 tokens did not affect the results substantially, so we report the results from the full set of transcripts.

following Kyle and Crossley (2015). A lower score on lexical frequency indicates that a transcript contains more rare words.

*Contextual diversity.* Another measure of lexical sophistication reflects contextual diversity, that is, the proportion of different documents in which a word token appears. As with lexical frequency, we obtained lexical contextual diversity values from the SUBTLEX-US and EsPal reference subtitle corpora, calculating the mean of the log contextual diversities for all tokens in each transcript. This measure has been demonstrated to contribute unique variance to a number of psycholinguistic constructs (e.g., Adelman et al., 2006; Baese-Berk et al., 2021; Kyle & Crossley, 2016). A lower score on contextual diversity indicates that a transcript contains words used in more unique contexts.

*Mean length of utterance in words.* We calculated the mean length of utterance in words (MLU-w) for each transcript, a commonly used measure of grammatical complexity (e.g., Ece Demir-Lira et al., 2019). We chose to compute this measure in words, rather than morphemes, to facilitate comparison across the English and Spanish transcripts (Gutiérrez-Clellen et al., 2000). A higher MLU-w indicates more words per utterance on average.

*Proportion of complex utterances.* The proportion of complex utterances indicates the number of utterances containing at least two lexical verbs divided by the total number of utterances (following Cameron-Faulkner & Noble, 2013). A higher proportion of complex utterances in a transcript reflects a higher level of syntactic complexity.

### ***Modeling***

We then fit a linear mixed effects model with the value of a measure as the outcome variable, and source and language group as predictors, along with the interaction between source and language group. We also fitted random effects of transcripts nested within families. For each of our six measures, we excluded extreme values, defined as values over 3.5 interquartile ranges away from the median; extreme values constituted < 4% of values for all metrics except for MTLTD, for which they constituted 6.1% of values for English and 5.4% of values for Spanish. Source was coded with sum contrasts, with book text as the reference level. The full specification

of each model was as follows:  $\text{value} \sim \text{source} * \text{language} + (1 | \text{family\_id/transcript\_id})$ . Using the fitted models, we computed pairwise comparisons among estimated marginal means, using Bonferroni corrections for multiple comparisons. If the model indicated significant interaction effects between source and language, we included comparisons across sources by language; if there were no interaction effects between source and language, we only included comparisons across sources.

## Results: Analysis 2

Table 3 provides an overview of the descriptive statistics for the six metrics of richness by source and language group.

**Table 3**

*Overview of descriptive statistics for lexical and grammatical metrics (median, mean absolute deviation, and range) by source for the English- ( $n = 22$ ) and Spanish-speaking ( $n = 20$ ) families.*

	English				Spanish			
	Book text	Read-aloud	Spont-book	Spont-other	Book text	Read-aloud	Spont-book	Spont-other
<i>N</i> families	21	22	22	22	15	18	20	20
<i>N</i> transcripts	53	42	48	235	30	33	37	198
MTLD	50.38 (22.49, 3.75–302.46)	36.97 (12.43, 11.09–100.07)	31.98 (7.34, 12.95–90.72)	34.34 (10.53, 6.11–123.48)	65.42 (32.24, 10.67–1088.64)	29.38 (16.59, 9.59–87.77)	27.19 (7.30, 12.68–56.46)	25.87 (7.94, 3.00–131.44)
Lexical density	0.40 (0.05, 0.04–0.89)	0.39 (0.04, 0.17–0.58)	0.29 (0.03, 0.17–0.59)	0.31 (0.04, 0.00–1.00)	0.42 (0.05, 0.28–0.94)	0.40 (0.05, 0.18–0.53)	0.32 (0.04, 0.08–0.44)	0.31 (0.04, 0.00–0.67)
Mean lexical frequency	2.89 (0.18, 1.56–3.26)	2.86 (0.13, 0.54–3.36)	3.10 (0.10, 2.13–3.41)	3.15 (0.10, -0.20–3.79)	2.69 (0.14, 0.64–3.21)	2.78 (0.09, 2.52–3.94)	2.99 (0.11, 2.47–3.22)	2.93 (0.13, 1.56–3.92)
Mean lexical contextual diversity	1.64 (0.11, 0.79–1.84)	1.62 (0.10, -0.01–1.86)	1.72 (0.06, 1.26–1.89)	1.75 (0.05, -0.81–2.00)	1.48 (0.11, 0.10–1.76)	1.54 (0.07, 1.40–1.84)	1.65 (0.05, 1.34–1.78)	1.65 (0.07, 0.81–1.99)
MLU-w	6.39 (1.37, 1.43–14.04)	6.02 (1.27, 1.20–10.86)	3.75 (0.48, 1.51–5.38)	3.58 (0.56, 1.00–7.75)	5.38 (1.69, 1.15–12.59)	4.38 (1.31, 2.35–9.68)	2.93 (0.62, 1.73–5.71)	2.82 (0.61, 1.00–5.90)
Proportion of complex utterances	0.19 (0.11, 0.00–0.97)	0.16 (0.08, 0.00–0.70)	0.04 (0.03, 0.00–0.24)	0.04 (0.04, 0.00–0.50)	0.10 (0.07, 0.00–0.59)	0.08 (0.08, 0.00–0.35)	0.02 (0.02, 0.00–0.18)	0.01 (0.01, 0.00–0.40)

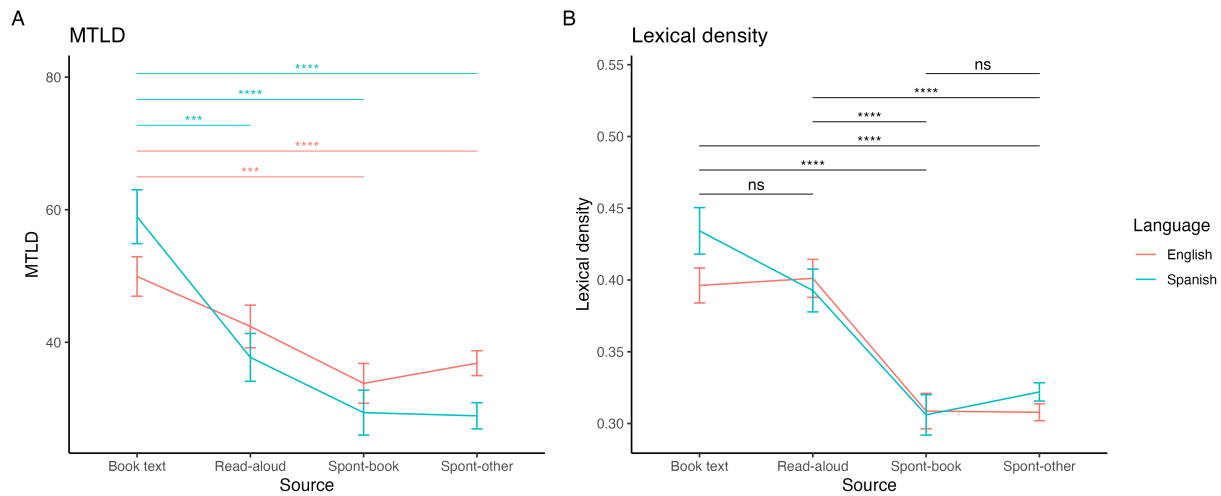
## Lexical diversity and density

Figure 3A shows the results for lexical diversity (i.e., MTLD). Analyses indicated a main effect of source such that MTLD was significantly higher in book text than spontaneous-book ( $p_{\text{EN}} < .001$ ,  $p_{\text{ES}} < .001$ ) and spontaneous-other ( $p_{\text{EN}} < .001$ ,  $p_{\text{ES}} < .001$ ) in both languages. However, there was a significant interaction between source and language, such that book text had

significantly higher MTLD than read-aloud in Spanish ( $p_{ES} < .001$ ), but not English ( $p_{EN} = .315$ ). Note also that although read-aloud text had numerically higher MTLD than both spontaneous-book and spontaneous-other speech, this difference did not reach significance (all  $ps > .090$ ). Spontaneous-book and spontaneous-other speech also did not significantly differ in MTLD (all  $ps > .999$ ). Thus, in this measure, both types of spontaneous speech tended to have lower MTLD values than book text.

### Figure 3

*Estimated marginal means for (A) lexical diversity (MTLD) and (B) lexical density as a function of source and language group. Error bars reflect standard errors. Non-significant contrasts in (A) have been omitted for presentational clarity.*



\*:  $p < .05$ , \*\*:  $p < .01$ , \*\*\*:  $p < .001$ , \*\*\*\*:  $p < .0001$

Results for lexical density shown in Figure 3B revealed significant main effects of source but not language group and no significant interaction between source and language group. Overall, book text and read-aloud text had significantly higher lexical density than spontaneous-book and spontaneous-other (all  $ps < .001$ ). Book text did not significantly differ from read-aloud text ( $p > .999$ ) and spontaneous tCDS did not differ in lexical density whether that speech occurred during book-sharing rather than during other types of activities ( $p > .999$ ). Thus, in this measure, we saw the consistent pattern that the two book-text sources patterned



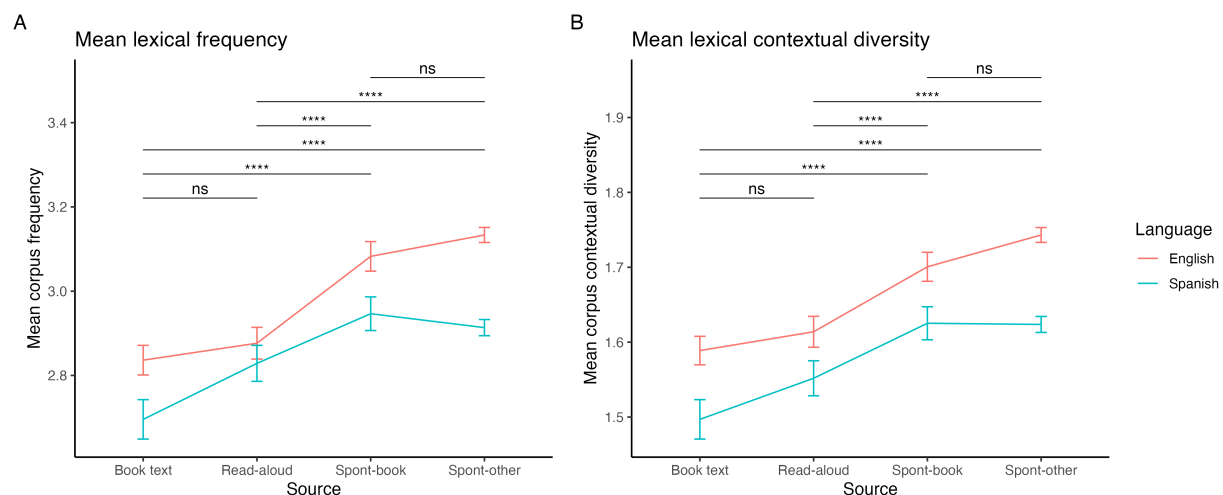
similarly and showed consistently more lexical density than both sources of spontaneous speech.

***Lexical sophistication (lexical frequency and contextual diversity)***

As shown in Figure 4A, lexical frequency was significantly different as a function of both source and language, but there was no significant interaction between source and language. Overall, book text and read-aloud text had significantly lower mean lexical frequency values than both spontaneous-book and spontaneous-other (all  $p$ s < .001), suggesting that book texts were likely to contain more rare words. Lexical frequency values did not significantly differ between book text and read-aloud text ( $p = .218$ ) and spontaneous tCDS did not differ in book-sharing and non-book-sharing contexts ( $p > .999$ ). There was also a main effect of language group, such that the transcripts from the English-speaking families had lexical frequency values that were higher than those of the Spanish-speaking families ( $b = 0.14$  [0.08, 0.19],  $p < .001$ ). This may reflect a real difference or could be an artifact of the estimates of lexical frequency that were used.

**Figure 4**

*Estimated marginal means for (A) mean lexical frequency and (B) contextual diversity and as a function of source and language group. Error bars reflect standard errors.*



\*:  $p < .05$ , \*\*:  $p < .01$ , \*\*\*:  $p < .001$ , \*\*\*\*:  $p < .0001$

Figure 4B also shows significant main effects for mean lexical contextual diversity of source and language group, but no significant interaction effects between source and language

group. Overall, book text and read-aloud text had significantly lower mean token contextual diversity than spontaneous-book and spontaneous-other (all  $p$ s < .001), while book text did not differ from read-aloud ( $p = .394$ ) and spontaneous tCDS did not differ across book-sharing vs. non-book-sharing contexts ( $p = .754$ ). Again, mean lexical contextual diversity values were higher overall in the transcripts from the English- compared to the Spanish-speaking families ( $b = 0.09 [0.05, 0.12]$ ,  $p < .001$ ), which could be an artifact of the estimates of lexical contextual diversity used.

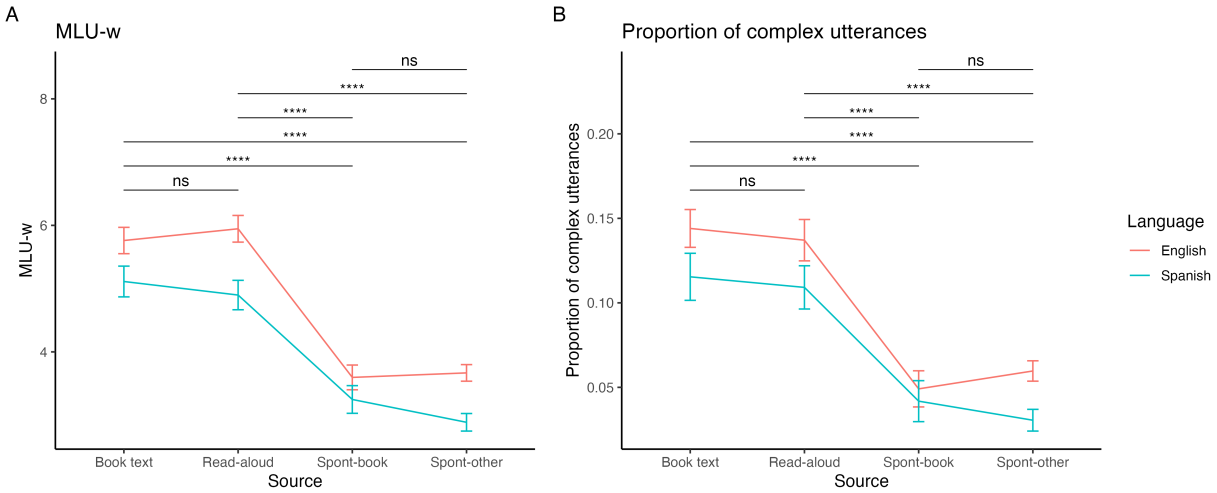
### *Grammatical measures*

Results for MLU-w are shown in Figure 5A. The model for MLU-w suggested significant main effects of source and language group and no significant interaction effects between source and language group. Overall, book text and read-aloud texts had significantly higher MLU-w than spontaneous-book and spontaneous-other (all  $p$ s < .001). MLU-w values did not significantly differ in book text and read-aloud ( $p > .999$ ), and values were not different in spontaneous tCDS produced during book-sharing vs. non-book-sharing contexts ( $p > .999$ ). Values were generally higher for the English- compared to the Spanish-speaking families ( $b = 0.71 [0.31, 1.10]$ ,  $p < .001$ ). These estimates may be greater in English due to an artifact of differences in how this measure is computed in morphologically-simpler languages like English, compared to morphologically-complex languages like Spanish (Gutiérrez-Clellen et al., 2000).

A similar pattern is observed in our final measure, proportion of complex utterances, in Figure 5B. That is, we see significant main effects of source and language, and no significant interaction between source and language. Overall, book text and read-aloud texts had a significantly greater proportion of complex utterances than spontaneous-book and spontaneous-other (all  $p$ s < .001), while book text did not differ from read-aloud ( $p > .999$ ) and the two sources of spontaneous tCDS did not differ from each other ( $p > .999$ ). Again, English values were generally higher than those from the Spanish ( $b = 0.02 [0.00, 0.04]$ ,  $p = .017$ ), suggesting more complex utterances produced by the English- compared to the Spanish-speaking families.

**Figure 5**

*Estimated marginal means for (A) mean length of utterances in words (MLU-w) and (B) proportion of complex utterances as a function of source and language group. Error bars reflect standard errors.*



∗:  $p < .05$ , ∗∗:  $p < .01$ , ∗∗∗:  $p < .001$ , ∗∗∗∗:  $p < .0001$

### ***Why did language group moderate the effects of MTLT?***

To summarize, we found that book text and read-aloud tCDS tended to pattern together, and spontaneous tCDS in book-sharing and non-book-sharing contexts tended to pattern together. These results were broadly consistent across measures, although there was an surprising source-by-language interaction for MTLT, such that book text had significantly higher MTLT than read-aloud in Spanish, but not English. This finding was unexpected, as it implied that what was read aloud in our Spanish-speaking families differed from what was in the book texts themselves. We thus explored potential moderating factors that may explain this difference. It is important to emphasize that these analyses are exploratory, and are likely underpowered, but they may provide suggestions for future confirmatory research.

First, we posited that the language in which the book texts were written may have affected MTLT. English-speaking caregivers only read English-language books to their children, whereas Spanish-speaking caregivers read books that were in one or two languages, sometimes in the same

book, i.e., English-only, Spanish-only, and bilingual books. Given that caregivers tended to be primarily monolingual Spanish speakers, they may have read more Spanish than English across the available book texts. It is difficult to directly compare across these different book types since bilingual books are likely to have higher MTLD than monolingual books, by definition, because there are two different languages from which to draw lexical tokens.

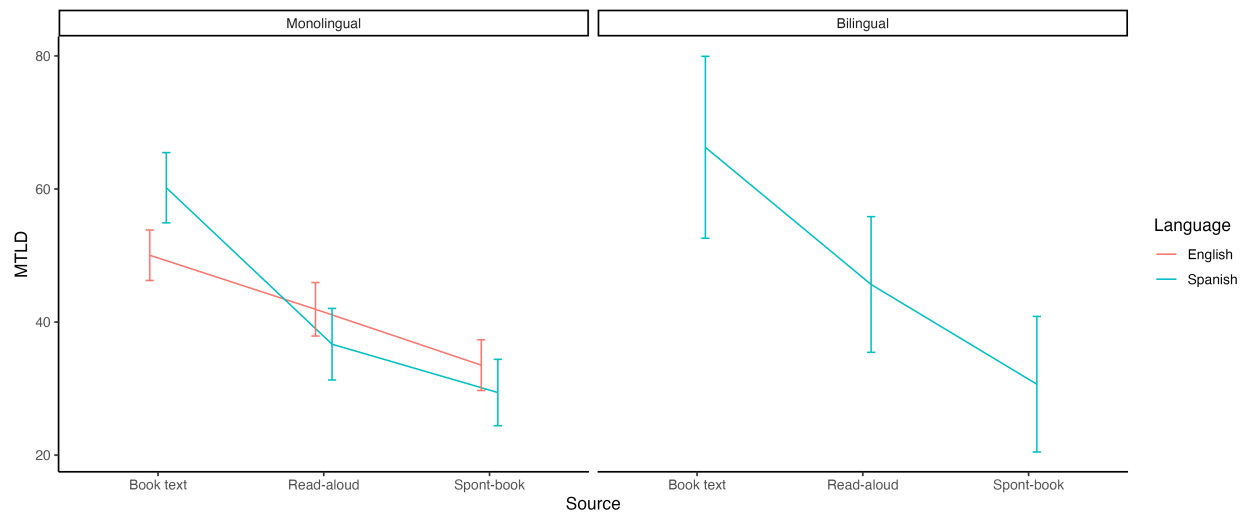
To address this issue, we split the book texts into monolingual and bilingual books, coding the read-aloud and spontaneous-book tCDS in terms of the proportions of books read within that transcript that were bilingual (i.e., if 1 monolingual and 3 bilingual books were read within a single transcript, then that transcript would be coded as 75% bilingual). We then ran a linear mixed-effects model with MTLD as the outcome variable, and source, language group, and bilingual proportion as predictors, along with all interactions. The estimated marginal means are shown in Figure 6. Including the proportion of the transcripts that contained bilingual books as a fixed effect did not eliminate the interaction between source and language group ( $p = .041$ ). Post-hoc analyses further indicated that when considering model estimates for monolingual books in the Spanish-speaking families, book text had significantly higher MTLDs than read-aloud text ( $p_{ES} = .007$ ), whereas this difference was not significant for English ( $p_{EN} > .999$ ).

We next posited that MTLD could be affected by the type of books that were read. Books written in prose are less lexically dense than non-prose books (e.g., rhyming books and naming books; Dawson et al., 2021), which may lead to greater MTLD in non-prose books. Books for Spanish-speaking families were more often non-prose books ( $23/30 = .77$ ) than the English-speaking families ( $27/55 = .49$ ). If reading strategies for non-prose books differed from strategies for prose books, e.g., incorporating more repetition for non-prose books thereby lowering MTLD, we would see more influence of book type in Spanish- than English-speaking families, thereby accounting for the observed source-by-language group interaction.

To explore this possibility, we split the book texts into non-prose and prose books, and coded the read-aloud and spontaneous-book tCDS in terms of the proportions of books read within that transcript that were prose. We then ran a linear mixed-effects model with MTLD as

**Figure 6**

*Estimated marginal means for MTLD for transcripts from different sources by book language. Error bars reflect standard errors.*



the outcome variable, and source, language group, and prose proportion as predictors, along with all interactions. The estimated marginal means are shown in Figure 7. The pattern of results was more similar across language groups for prose books, with spontaneous-book speech having lower MTLD than book text for families in both language groups ( $p_{EN} < .001$ ,  $p_{ES} = .047$ ).

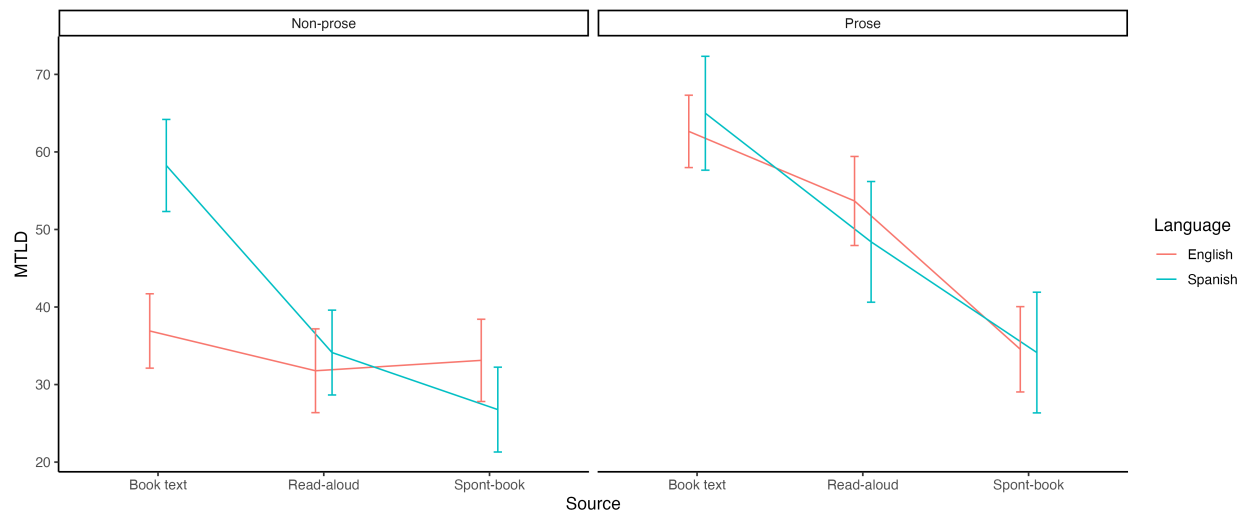
Spontaneous-book text had numerically lower MTLD than read-aloud, although this difference was only significant in English ( $p_{EN} = .025$ ,  $p_{ES} > .999$ ). However, for non-prose books, book text had significantly greater MTLD than read-aloud ( $p_{ES} = .016$ ) and spontaneous-book ( $p_{ES} < .001$ ) in Spanish, but not in English (both  $ps > .999$ ). This suggests that differing proportions of non-narrative books in the two samples of families may explain some of the observed variance in MTLD and the source-by-language interaction.

## Discussion: Analysis 2

Taken together, these analyses showed that book texts and read-aloud speech (speech actually read from those book texts) tended to have greater lexical diversity, higher proportions of content words, contain more rare and distinctive vocabulary words, and be comprised of longer and more morphosyntactically complex sentences than speech that is spoken spontaneously

**Figure 7**

*Estimated marginal means for MTLD for transcripts from different sources by book type. Error bars reflect standard errors.*



during everyday conversations with young children. Importantly, differences in the types of books (e.g., prose vs. non-prose) that families read could account for some of the observed variance in MTLD across families from different language backgrounds.

### General discussion

In this study, we investigated how sharing books can influence caregiver speech to 2-year-old children. We conducted two types of analyses to explore the linguistic characteristics of different language sources available to young children during everyday activities. The first set of analyses used a “mega transcript” approach that allowed us to model overall trends in lexical diversity across different sources of language, using number of types and the TTR, two measures that have been at the forefront of many previous analyses of linguistic richness (Dawson et al., 2021; Montag et al., 2015). In our second set of analyses, we tested patterns that held across individuals and expanded our view to include six different measures of lexical and grammatical “richness” in speech. In both analyses, we compared across four sources of language (book text, read aloud text, spontaneous tCDS during book sharing, and spontaneous tCDS in other contexts) and between the two language groups (families from English-speaking and Spanish-speaking

households). Notably, these two analytic approaches both showed that the distinct “richness” in caregiver speech during book sharing was largely driven by the book text itself. Thus, our findings provided further evidence that written texts are a rich source of lexical and grammatical information that may be particularly valuable for language learning.

Further, we examined the text from books that were specifically available in the homes of the families in our studies—in particular, those of English- and Spanish-speaking families in the western United States. This work is in contrast to prior studies that have conducted similar analyses of texts from curated book lists of popular or best-selling titles for young children. In fact, of the 86 books that were spontaneously read by our families, only 5 were in the 100-book American English corpus analyzed previously by Montag et al. (2015). In spite of the different samples of books, our results aligned with these prior studies, such that the text of these books read in the home, both in their entirety and what was read aloud, was richer compared to the spontaneous speech directed to young children during book sharing and other activities.

We found that the richness of book texts and read-aloud text patterned together on most measures, suggesting that caregivers were reading directly from the texts. While this may not seem surprising at first, this could only be true if, on the whole, caregivers actually read what was in the texts of the books, rather than truncating or simplifying the text of the books. This is an important finding because while the books in the home may be lexically- and grammatically-rich sources of language for children, if caregivers did not stay true to those texts during actual reading (either consciously or unconsciously), their contribution to the linguistic environment of the child could have been diminished. We also found that the features of spontaneous tCDS during book shared patterns with spontaneous tCDS in non-book sharing contexts on most measures. Thus, when caregivers talk with children spontaneously, the features of the tCDS are not necessarily lexically or grammatically richer when it occurs during a book reading session than during other types of activities.

In general, results were similar across the two language groups (English and Spanish) that we sampled. This is especially exciting given that our sample of families represented two

prominent language groups in the U.S., expanding on previous work representing only monolingual American and British English. One exception to the similarities between the two language groups was that there was a greater difference in the lexical diversity of book texts versus read-aloud text in our Spanish- compared to our English-speaking families. Interestingly, this pattern of findings remained even when we restricted our analyses to books that were written in a single language, rather than bilingual books that, by their nature, are likely to be more lexically diverse. Separating the books into prose and non-prose genres, however, did allow us to account for some of the observed variance between languages in terms of lexical diversity. Non-prose books (e.g., books focused on vocabulary, or nursery rhymes) had greater differences in lexical diversities between the two languages, compared to prose books. This set of findings may have been impacted by our relatively small sample size. Thus, future research should further explore the potential sources of cross-linguistic differences in lexical diversity using larger samples of books and families.

Overall, we found that book text and read aloud text were lexically and grammatically richer than spontaneous tCDS on all measures. Thus, the richness of language during book sharing is largely driven by the characteristics of the book texts that are read aloud, rather than by the spontaneous speech that also occurs during book sharing. Read-aloud text is just one part of the language experienced by children in these book-sharing activities. As caregivers read a book aloud, the accompanying spontaneous speech, while not as lexically or grammatically rich as the text in the books, may nevertheless serve other important functions. For example, a caregiver may read aloud a multi-clause sentence with a sophisticated vocabulary word and then stop to say, “yeah, you see,” or respond to a child’s repetition of the word with “that’s right.” Such extratextual speech may help to affirm children’s understanding, promote the child’s engagement with the book, highlight the rich language, and prompt dialogue, which in turn may support children’s learning (e.g., [Blewitt & Langan, 2016](#); [Mol et al., 2008](#); [Read et al., 2023](#)).



## Limitations

Our measures focused specifically on particular operationalizations of lexical and grammatical diversity which may have reduced or inflated our effects. For example, language group differences in MTLTD may have been especially sensitive to the types of books that families read given that “vocabulary” books with just a few words on a page yield a less stable measure of linguistic diversity than other types of books. Moreover, our measures may not have captured important dimensions of high-quality language that supports children’s learning. We did not evaluate the speech produced by the target children during shared book reading, or features of the back-and-forth conversational turns that occurred in the spontaneous talk which have been found to correlate with language learning (e.g., [Gilkerson et al., 2017](#)). Thus, we have not counted the dialogic nature of the spontaneous speech in our definition of richness. In addition, we have not categorized or measured the content of the speech (e.g., the degree to which extratextual talk is related to the book or how much it is decontextualized or cognitively demanding, e.g., [Read et al., 2023](#)). Future research should continue to broaden the ways in which rich and supportive language is measured in order to better capture the added value of shared book reading to early language development.

Future research should also investigate the relation between the types of books that are read and the variation in richness across language sources. We perceive it as a strength that the corpora of children’s book texts used here represented the books available in the homes of our participants, thereby increasing the ecological validity of our corpora. However, we could not control book type across families or language groups. Previous research has demonstrated impacts on the amount of spontaneous talk in book sharing episodes based on whether the books themselves were considered “high text” vs. “low text” or whether books are considered “informative” vs. “narrative” in genre ([J. Anderson et al., 2004](#); [Fletcher & Finch, 2014](#); [Price et al., 2009](#); [Read et al., 2023](#)) often revealing an inverse relationship wherein less book text induces more spontaneous speech. Investigating the impacts of book type on measures of richness more fully, with larger naturally selected corpora of books chosen and read in the home, could further

deepen our understanding of the benefits of shared reading for young children's language learning.

Finally, we sampled only a few of the shared reading episodes that were likely to have occurred in our families. Our analyses were based on day-long audio recordings, and were limited to families with 2-year-old children. The books that caregivers read and the types of talk that they use while reading are likely to become more lexically diverse, sophisticated, and grammatically complex as children progress from toddlerhood through preschool. Research has demonstrated that the amount and content of extratextual book sharing talk also varies with child age and ability (McArthur et al., 2005; Peralta de Mendoza, 1995; Wheeler, 1983). Thus, further research should explore how the linguistic richness of caregiver speech might change along with a child's development.

### **Conclusion**

The results from this study illustrated several ways in which words on the page have tremendous power for exposing children to rich language. Using day-long recordings, we explored the lexical and grammatical features of speech produced in naturally-occurring book-sharing episodes and other activities with young children. In two quantitative analyses and across two language groups, we extended earlier findings that speech based on book texts was more lexically diverse and grammatically complex than spontaneous child-directed speech. These findings highlight that linguistic richness varies both within and across activities and provide a strong empirical basis for understanding how book sharing positively impacts child language outcomes.

## References

- Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, 17(9), 814–823. <https://doi.org/10.1111/j.1467-9280.2006.01787.x>
- Anderson, J., Anderson, A., Lynch, J., & Shapiro, J. (2004). Examining the effects of gender and genre on interactions in shared book reading. *Reading Research and Instruction*, 43(4), 1–20. <https://doi.org/10.1080/19388070409558414>
- Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & Madigan, S. (2021). Linking Quality and Quantity of Parental Linguistic Input to Child Language Skills: A Meta-Analysis. *Child Development*, 92(2), 484–501. <https://doi.org/10.1111/cdev.13508>
- Arterberry, M. E., Bornstein, M. H., Midgett, C., Putnick, D. L., & Bornstein, M. H. (2007). Early attention and literacy experiences predict adaptive communication. *First Language*, 27(2), 175–189. <https://doi.org/10.1177/0142723706075784>
- Baese-Berk, M. M., Drake, S., Foster, K., Lee, D., Staggs, C., & Wright, J. M. (2021). Lexical Diversity, Lexical Sophistication, and Predictability for Speech in Multiple Listening Conditions. *Frontiers in Psychology*, 12. <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.661415>
- Bang, J. Y., Mora, A., Munévar, M., Fernald, A., & Marchman, V. A. (2022, September 29). *Time to talk: Multiple sources of variability in caregiver verbal engagement during everyday activities in English- and Spanish-speaking families in the U.S.* <https://doi.org/10.31234/osf.io/6jzww>
- Berman, R., & Nir, B. (2010). The lexicon in writing–speech-differentiation. *Written Language & Literacy*, 13(2), 183–205. <https://doi.org/10.1075/wll.13.2.01ber>
- Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12), 1153–1170. <https://doi.org/10.1016/j.tics.2022.09.015>
- Blewitt, P., & Langan, R. (2016). Learning words during shared book reading: The role of

- extratextual talk designed to increase child engagement. *Journal of Experimental Child Psychology*, 150, 404–410. <https://doi.org/10.1016/j.jecp.2016.06.009>
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. <https://doi.org/10.3758/BRM.41.4.977>
- 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>
- 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*, 1(1). <https://doi.org/10.34842/5WE1-YK94>
- Debaryshe, B. D. (1993). Joint picture-book reading correlates of early oral language skill. *Journal of Child Language*, 20(2), 455–461. <https://doi.org/10.1017/S0305000900008370>
- Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, 45(4), 1246–1258. <https://doi.org/10.3758/s13428-013-0326-1>
- Ece Demir-Lira, Ö., Applebaum, L. R., Goldin-Meadow, S., & Levine, S. C. (2019). Parents’ early book reading to children: Relation to children’s later language and literacy outcomes controlling for other parent language input. *Developmental Science*, 22(3), e12764. <https://doi.org/10.1111/desc.12764>
- Farrant, B. M., & Zubrick, S. R. (2012). Early vocabulary development: The importance of joint attention and parent-child book reading. *First Language*, 32(3), 343–364. <https://doi.org/10.1177/0142723711422626>
- Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects. *Journal of Speech, Language, and Hearing Research*, 58(3), 840–852. [https://doi.org/10.1044/2015\\_JSLHR-L-14-0280](https://doi.org/10.1044/2015_JSLHR-L-14-0280)
- Flack, Z. M., Field, A. P., & Horst, J. S. (2018). The effects of shared storybook reading on word

- learning: A meta-analysis. *Developmental Psychology*, 54(7), 1334–1346.  
<https://doi.org/10.1037/dev0000512>
- Fletcher, K. L., & Finch, W. H. (2014). The role of book familiarity and book type on mothers' reading strategies and toddlers' responsiveness. *Journal of Early Childhood Literacy*.  
<https://doi.org/10.1177/1468798414523026>
- Fletcher, K. L., & Reese, E. (2005). Picture book reading with young children: A conceptual framework. *Developmental Review*, 25(1), 64–103. <https://doi.org/10.1016/j.dr.2004.08.009>
- Gilkerson, J., & Richards, J. A. (2008). *The LENA Natural Language Study*.
- Gilkerson, J., Richards, J. A., & Topping, K. J. (2017). The impact of book reading in the early years on parent–child language interaction. *Journal of Early Childhood Literacy*, 17(1), 92–110. <https://doi.org/10.1177/1468798415608907>
- Gutiérrez-Clellen, V. F., Restrepo, M. A., Bedore, L., Peña, E., & Anderson, R. (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. *Language, Speech, and Hearing Services in Schools*, 31(1), 88–98.  
<https://doi.org/10.1044/0161-1461.3101.88>
- Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample Size and Type-Token Ratios for Oral Language of Preschool Children. *Journal of Speech, Language, and Hearing Research*, 29(1), 129–134. <https://doi.org/10.1044/jslr.2901.129>
- Hindman, A. H., Skibbe, L. E., & Foster, T. D. (2014). Exploring the variety of parental talk during shared book reading and its contributions to preschool language and literacy: Evidence from the Early Childhood Longitudinal Study-Birth Cohort. *Reading and Writing*, 27(2), 287–313. <https://doi.org/10.1007/s11145-013-9445-4>
- Hoff-Ginsberg, E. (1991). Mother-Child Conversation in Different Social Classes and Communicative Settings. *Child Development*, 62(4), 782–796.  
<https://doi.org/10.2307/1131177>
- Karrass, J., & Braungart-Rieker, J. M. (2005). Effects of shared parent–infant book reading on early language acquisition. *Journal of Applied Developmental Psychology*, 26(2), 133–148.

<https://doi.org/10.1016/j.appdev.2004.12.003>

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, 01(1), 60–69.

<https://doi.org/10.7820/vli.v01.1.koizumi>

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12–24.

<https://doi.org/10.1016/j.jslw.2016.10.003>

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, 49(4), 757–786.

<https://doi.org/10.1002/tesq.194>

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. (3rd ed.). Lawrence Erlbaum Associates.

Massaro, D. W. (2015). Two Different Communication Genres and Implications for Vocabulary Development and Learning to Read. *Journal of Literacy Research*, 47(4), 505–527.

<https://doi.org/10.1177/1086296X15627528>

McArthur, D., Adamson, L., & Deckner, D. F. (2005). As Stories Become Familiar: Mother-Child Conversations During Shared Reading. *Merrill-Palmer Quarterly*, 51(4), 389–411.

<https://muse.jhu.edu/pub/27/article/189375>

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* [PhD thesis, The University of Memphis].

<https://www.proquest.com/docview/305349212/abstract/AB75BC58B2FF43BEPQ/1>

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. <https://doi.org/10.3758/BRM.42.2.381>

Michalke, M. (2021). *koRpus: Text analysis with emphasis on POS tagging, readability, and lexical diversity* [Manual]. <https://reaktanz.de/?c=hacking&s=koRpus>

- 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. (2019). Differences in sentence complexity in the text of children’s picture books and child-directed speech. *First Language*, 39(5), 527–546. <https://doi.org/10.1177/0142723719849996>
- 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>
- Noble, C. H., Cameron-Faulkner, T., & Lieven, E. (2018). Keeping it simple: The grammatical properties of shared book reading. *Journal of Child Language*, 45(3), 753–766. <https://doi.org/10.1017/S0305000917000447>
- Peralta de Mendoza, O. A. (1995). Developmental Changes and Socioeconomic Differences in Mother-Infant Picturebook Reading. *European Journal of Psychology of Education*, 10(3), 261–272. <https://www.jstor.org/stable/23420013>
- Price, L. H., Van Kleeck, A., & Huberty, C. J. (2009). Talk During Book Sharing Between Parents and Preschool Children: A Comparison Between Storybook and Expository Book Conditions. *Reading Research Quarterly*, 44(2), 171–194. <https://doi.org/10.1598/RRQ.44.2.4>
- Raikes, H., Alexander Pan, B., Luze, G., Tamis-LeMonda, C. S., Brooks-Gunn, J., Constantine, J., Banks Tarullo, L., Abigail Raikes, H., & Rodriguez, E. T. (2006). Mother–Child Bookreading in Low-Income Families: Correlates and Outcomes During the First Three Years of Life. *Child Development*, 77(4), 924–953. <https://doi.org/10.1111/j.1467-8624.2006.00911.x>
- Read, K., Contreras, P., & Martinez, H. (2021). Tres formas: Shared reading practices with three types of Spanish and English dual-language learning preschoolers. *Bilingual Research Journal*, 44(3), 360–380. <https://doi.org/10.1080/15235882.2021.1994485>
- Read, K., Rabinowitz, S., & Harrison, H. (2023). It’s the talk that counts: A review of how the extra-textual talk of caregivers during shared book reading with young children has been

- categorized and measured. *Journal of Early Childhood Literacy*, 14687984231202968.  
<https://doi.org/10.1177/14687984231202968>
- Sénéchal, M., & LeFevre, J.-A. (2014). Continuity and Change in the Home Literacy Environment as Predictors of Growth in Vocabulary and Reading. *Child Development*, 85(4), 1552–1568. <https://doi.org/10.1111/cdev.12222>
- Sénéchal, M., LeFevre, J.-A., Hudson, E., & Lawson, E. P. (1996). Knowledge of storybooks as a predictor of young children's vocabulary. *Journal of Educational Psychology*, 88(3), 520–536. <https://doi.org/10.1037/0022-0663.88.3.520>
- Sénéchal, M., Pagan, S., Lever, R., & Ouellette, G. P. (2008). Relations Among the Frequency of Shared Reading and 4-Year-Old Children's Vocabulary, Morphological and Syntax Comprehension, and Narrative Skills. *Early Education and Development*, 19(1), 27–44. <https://doi.org/10.1080/10409280701838710>
- Singh, L., & Rajendra, S. J. (2024). Greater attention to socioeconomic status in developmental research can improve the external validity, generalizability, and replicability of developmental science. *Developmental Science*, 27(5), e13521. <https://doi.org/10.1111/desc.13521>
- Soderstrom, M., & Wittebolle, K. (2013). When Do Caregivers Talk? The Influences of Activity and Time of Day on Caregiver Speech and Child Vocalizations in Two Childcare Environments. *PLOS ONE*, 8(11), e80646. <https://doi.org/10.1371/journal.pone.0080646>
- Stills, M. (2016). *Language Sample Length Effects on Various Lexical Diversity Measures: An Analysis of Spanish Language Samples from Children* [Honor's thesis, Portland State University]. <https://pdxscholar.library.pdx.edu/honorstheses/233>
- Stoops, A., & Montag, J. L. (2024). A novel corpus of naturalistic picture book reading with 2-to-3 year old children. *Language Development Research*, 4(1, 1). <https://doi.org/10.34842/3kz6-4s17>
- Strömquist, S., Johansson, V., Kriz, S., Ragnarsdóttir, H., Aisenman, R., & Ravid, D. (2002). Toward a cross-linguistic comparison of lexical quanta in speech and writing. *Written Language & Literacy*, 5(1), 45–67. <https://doi.org/10.1075/wll.5.1.03str>



- Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., & Lo, T. (2019). Routine Language: Speech Directed to Infants During Home Activities. *Child Development*, 90(6), 2135–2152. <https://doi.org/10.1111/cdev.13089>
- Weizman, Z. O., & Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*, 37(2), 265–279. <https://doi.org/10.1037/0012-1649.37.2.265>
- Wheeler, M. P. (1983). Context-related age changes in mothers' speech: Joint book reading. *Journal of Child Language*, 10(1), 259–263. <https://doi.org/10.1017/s0305000900005304>