Mining Big Data to Analyze Influences on Children's Language Acquisition

Arabella Williams, Elma Tudjinovic, Toby Baylon, Jessica Alarcon, Rukmini Bose, and Dominic W. Massaro

University of California, Santa Cruz
Santa Cruz, CA 95060

Acknowledgement
This research was support in part by a 2018-19 Dickson Emeriti Professorship Award to Dom Massaro. Special thanks to E. H. Hiebert for discussions and for providing resources for the vocabulary analyses.

Abstract

Children are highly dependent on and influenced by the language they hear. Parental linguistic input has been repeatedly demonstrated as one of the primary determinants of a child's language acquisition. Parental and caretaker input includes reading aloud to children, which is additionally valuable because the vocabulary and grammar in picture books is usually richer and more extensive than that found in interactive speech. Previous research has shown, for example, that the vocabulary found in picture books features more relatively rare words and a higher linguistic complexity (as measured by accepted reading grade level measures) than found in speech between adults and children. In the present study, we replicate and extend this analysis to TV Media. Using a framework of formal versus informal language, TV Media includes formal as well as informal scripts that might place it somewhere between the language found in child-directed speech and picture books. A new database of parental speech was tested along with a variety of scripts from children's popular TV media. We found that picture books had more rare words than TV Media and parent's speech to children. In a new analysis, picture books also provided many more rare words that have been found important for school curricula. Seven measures of reading grade level revealed that picture books maintained a strong advantage over TV Media with an average grade level of 3.5 for picture books, 2.5 for TV Media, and 2.0 for child-directed speech. Thus, picture books maintain an advantage over TV Media in both vocabulary and linguistic complexity. TV Media could be valuable for children's cognitive and linguistic growth but it lacks the potential interactive component of child-directed speech and the cognitive and linguistic richness of picture books.

Introduction

As empirical examinations of language development have increased, a number of developmental trends have emerged in the literature. In particular, psycholinguistic research has revealed several salient facets of language development; including important findings regarding the qualities and influences of parental input, various linguistic principles, speech comprehension in infancy, adult-directed speech (ADS), child-directed speech (CDS), children's television media, and picture books. These findings are essential in accurately characterizing how language comprehension and production function in childhood and beyond. While children's word input comes primarily from their parents, children are often exposed to alternative forms of input, such as from picture books and TV (television-video) media.

Whitehurtst et al. (1988) examined how the form of engagement with picture books for children aged 21 to 35 months impacts their language learning. Parents in the experimental group were instructed to "increase their rates of open-ended questions, function/attribute, and expansions; to respond appropriately to children's attempts to answer these questions; and to decrease their frequency of straight reading and questions that could be answered by pointing while reading to their child over the course of a month (Whitehurst et al., 1998, p. 552). Compared to the control group who were simply instructed to read in their usual way, children in the experimental group scored higher on expressive language ability. Children in the experimental group also used more phrases, spoke fewer single words, and had a higher MLU (define?) than the control group. These findings suggest engaging children in dialog during reading picture books may encourage significant gains in expressive language acquisition (Whitehurst et al., 1988).

Evans and Saint-Aubin (2013) looked deeper into this link between vocabulary acquisition and picture books. By examining the eye movements of French preschoolers aged 50 to 62 months as they were read three picture books, they discovered that children's eye movements were stable and generally aimed toward the illustrations in the book (Evans & Saint-Aubin, 2013). Through administrations of a scale measure of children's receptive vocabulary (the ÉVIP) and adaptations of this scale measure, Evans and Saint-Aubin (2013) also found that children's receptive vocabulary of uncommon words increased over the course of the readings. Furthermore, it was revealed that this improvement was correlated with children's pretest receptive vocabulary (Evans & Saint-Aubin, 2013). Based on these findings, the authors concluded that "print affords children the opportunity to continually roam through the illustrations and search for matches between what is said and what is depicted, and to learn the meaning of new words" (Evans & Saint-Aubin, 2013, p. 607).

Shinksey (2020) examined how different forms of picture books may impact how children learn words. Using either a lift-the-flap book or a normal picture book, two-year-old children were taught a novel term for an unfamiliar food. When tested, children who were read the lift-the-flap book performed significantly worse in an identification task than children who saw the no-flap book (Shinskey, 2020). This finding was specific to the new words and did not generalize to children's recognition of higher frequency words in the books. The authors concluded that the outcome supports "cognitive load accounts suggesting that tactical features distract from the book's content" (Shinskey, 2020, p. 1), and may suggest that lift-the-flap books hinder word learning.

The influence of other children's media has been investigated on children's vocabulary acquisition, in particular, TV Media aimed toward babies. A study conducted by DeLoache et al. (2010) tested if infants aged 12 to 18 months would learn any words from a DVD targeted toward educating babies. This DVD had a runtime of 39 minutes and featured scenes of a house and a yard (DeLoache et al., 2010). A popular DVD which was advertised for ages 12 months and up showed footage of a house and yard which had voiced items around them. They found that after a month of exposure, the children who watched the DVD did not learn any more words than children who did not watch it (DeLoache et al., 2010). The most vocabulary gains actually occurred in an alternative group in which parents did not show any video to their child, and instead tried to teach their children the words themselves (DeLoache et al., 2010). The researchers concluded that what infants learn from baby media is minimal, and likely overestimated. This is consistent with previous findings that "very young children often fail to use information communicated to them via symbolic media, including pictures, models, and video" (DeLoache et al, 2010, p. 4). On the other hand, the child's linguistic input is easily envisioned as falling on a fuzzy continuum from live face-to-face dialog, remote video conferencing, audio phone calls, TV Media, to picture books. These interactions might have various degrees of effectiveness but it is unlikely that one of them would be totally ineffective.

Recently, several investigators established that the language in children's books is more complex and featured a more extensive vocabulary than child-directed speech and even adult-directed speech (Massaro, 2015a; 2017b). After eliminating the 5,000 most common English words from the Corpus of Contemporary American English (COCA, 1990-2012), a sampled database of children's picture books had three times as many rare words as a database of

CDS and 1.64 times as many rare words as a database of ADS (Massaro, 2015a; 2017b). One implication of this finding is that reading picture books aloud to children may expose them to a more extensive vocabulary at an earlier age than exposure to CDS or ADS alone.

It is also reasonable to measure linguistic complexity in terms of grade level readability. Readability can be best understood as the reading grade level required of a reader to effectively read some form of text (Readability Formulas, n.d.). There are several different formulas that have been used in calculating readability, each using factors such as the number of words, the length of sentences, and the average number of syllables to determine the reading grade level of the passage. These readability formulas are computed from each sentence as the basic unit of analysis. Other measures using the coherence across sentences in a passage such as Co-Metrix are not appropriate for child-directed speech. The readability measures (Readability Formulas, n.d.) used in the analysis are particularly appropriate for our comparisons of CDS, picture books, and TV Media because all three of the media switch topics to various degrees during recording sessions.

There were also substantial differences in the reading level of picture books, CDS, and ADS (Massaro, 2017b)). Reading grade level is accepted as a measure of linguistic complexity and it is reasonable to apply it to the transcribed text of spoken language. Massaro (2017b) used the Flesch-Kincaid grade level formula, Gunning Fog Index, Coleman grade level, SMOG index, and Automated Readability Index (ARI) to evaluate the readability of samples of CDS, ADS, and picture books. According to these measures of reading grade level, picture books had the highest average grade level (4.2) compared to both child-directed speech (1.9) and adult-directed speech (3.0). Thus, popular children's picture books had a higher reading difficulty than CDS

and ADS. This strengthens the argument that reading picture books to children is beneficial, as it appears to familiarize them with more challenging language and a more extensive vocabulary than they would otherwise encounter. Reading books to children thus provides a "linguistic and cognitive complexity not typically found in speech to children" (Massaro, 2017b, p. 63). Analogous to the importance of the frequency of CDS, it is important to note that the benefits of reading picture books aloud to children are contingent upon how frequently children are being read to.

The majority of research examining the quality of speech input focuses on child directed speech and adult directed speech. Only recently have investigators begun exploring the nature of the vocabulary and language used in children's books. It is important to note, however, that many children also spend a considerable amount of time watching television shows (Vandewater et al., 2006). A sobering observation is that children in America watch between 2 to 5 hours of television a day on average, which accounts for "more time [spent] than in any other single discretionary activity except for sleep" (Vanderwater et al., 2006, p. 2). Despite this, little empirical consideration has been placed on the linguistic content of children's television. Some observational research has associated television viewership, particularly time spent watching baby DVDs and videos every day, with lower scores on a measure of language development in children under two (Zimmerman et al., 2007). However, this research does not take into account the actual content of the media children are consuming, but rather speaks to the potential implications of TV viewership in early childhood. While this finding may point toward children's difficulty in learning from TV Media or whether it has a lower quality of linguistic content, research is needed to establish if this is the case. In the present study, we explored this

topic by sampling popular children's television shows and movies and comparing their vocabulary and computed reading grade levels to the reading grade levels of samples of picture books and CDS.

In the present investigation, we aim to evaluate TV Media in the same manner that we have analyzed CDS, ADS, and picture books. In addition, we replicate and extend our previous investigations. We refined one of our children's speech databases to include the age and sex of the children who are speaking. We also utilized a different database of CDS which included children aged 2 to 5, as compared to our previous study that used speech directed to 8-month-old children. It may be that the previous finding of a large advantage of picture books over CDS might have been due to the very young age of the children being spoken to, and not representative of a general linguistic advantage of picture books over child-directed speech.

Insert Table 1 About Here

Of special interest is our analysis of TV Media. Table 1 gives a taxonomy of the potential independence of language modality and formal (nonconversational) versus informal (conversational) dialogue. Table 1 illustrates examples of communication media in which the formality of the language and its modality are independent of one another. For example, a TED Talk could be spoken and formal whereas texting could be informal and written. Thus, we expect that the formal nature of picture books allows the writer to make more deliberate word and grammatical choices that are not possible in CDS. CDS requires a spontaneity (Grice, 1975) that necessarily limits word choice and allows deixis to substitute for various words and grammatical

constructions. We expect that the media in children's television programs might be intermediate between CDS and picture books. That is, TV Media is necessarily scripted but we might expect much of the script is aimed at spontaneous dialog resembling what is found in CDS.

Method

Database of Parent (CDS)

All parent and child speech data in the present study were derived from the Child Language Data Exchange System (CHILDES). The CHILDES system provides free access to thousands of transcriptions containing child and parent speech that have been contributed by researchers around the world. Our sample of parent speech consists of 145 transcripts uploaded to the CHILDES English-NA corpus by the following researchers and organizations: Bates, Bliss, Bloom, Bohannon, Braunwald, Brown, Clark, Demetras, EllisWeismer, Feldman, Garvey, Gelman, Gillam, Gleason, Haggerty, Hall, HSLLD, Kuczaj, MacWhinney, McCune, McMillan, Morisset, NewEngland, NewmanRatner, Peters, POLER, Post, Sachs, Snow, Suppes, TD, Valian, Van Houten, Van Kleeck, Warren, and Weist. Transcripts that included speech from children outside of our target age range (children aged 2-5) and transcripts that did not specify the gender of the participant were excluded from the sample. As the transcripts sampled for this study originate from a wide range of sources, parent speech data thus reflects a large variety of environments and situations. Some examples of transcript settings include meal/snack-time, storytelling, preschool, toy play, and free play.

As we were interested in whether parent speech differed across age of the children, sampled transcripts were separated into subsamples based on age (2, 3, 4, or 5). Given that we also included sex as a variable, eight subsamples in total were created, one dataset for each

possible gender/age pair (i.e., males aged 2, females aged 2, males aged 3, females aged 3, etc.). Utterances from these transcripts were extracted and counted using the computerized language analysis program CLAN by running the KWAL and MLU commands on each subsample with specifications for the target speaker. As many of these original utterances included linguistic information beyond just the words spoken by the participant (pitch change, tone of voice, non-verbal communication, etc.), utterances were edited to contain only transcripts of the spoken words.

Table 2 gives the number of transcripts and utterances for each sample and subsample. In total, the database includes 36,163 child utterances (56.26% male utterances and 43.76% female utterances) and 28,241 parent utterances (76.83% mother utterances and 23.17% father utterances) sourced from the CHILDES English-NA corpus. Due to there being significantly fewer utterances available in the corpus for children aged 5 than data for children aged 2-4, our sample consists of fewer utterances for 5-year-olds than 2-, 3-, and 4-year-olds. A similar pattern emerged for father's speech, with parent speech in the corpus primarily belonging to the mother. As such, our sample has more mother utterances (n=21,698) than father utterances (n=6,543). Furthermore, there were slightly more transcripts featuring male children (n=79) and utterances from male children (n=20,337) than transcripts featuring female children (n=66) and utterances from female children (n=15,826). The underrepresentation of older children, fathers, and female children in our sample may pose as a limitation, particularly in terms of applying or generalizing our findings to these groups.

Insert Table 2 About Here

For the readability analysis, our sample of CDS included six text samples of about 2,000 words collected from our database of parent speech, separated by the age and sex of the child featured in the transcript. Transcripts featuring children aged four and five were limited, so male and female children were combined in those samples to create a sufficient sample size. In order to ensure an accurate readability analysis, all sampled utterances were edited by hand to remove any non-speech information and revise any incorrect spelling, spacing, and punctuation.

Children's Picture Books

The picture book database was identical to that used in our earlier studies. The text from 112 popular picture books was transcribed for the analysis (Massaro, 2015b). Appendix 1 lists the books included in the database. The books were considered narrative as opposed to informational books.

For the reading grade level analysis, six text samples of approximately 2,000 words each were selected from the Massaro (2015b, 2017a) corpus containing the full text from 112 popular children's picture books. The number of samples and the size of each sample was chosen to match the CDS samples. Each sample contains the text of between one to ten full picture books combined into a single document.

TV Media

For the purpose of the present study, the TV Media text featured the spoken language from children's film and television scripts. These samples excluded stage directions, sound effects, instances of character names not directly uttered (i.e., denoting who is speaking in the

script), and credit texts. Given this editing, the sample only includes speech utterances that children watching the media will actually hear.

A few popular children's television shows, such as Sesame Street and Blue's Clues, were selected to create this database. Television show transcripts were primarily obtained from Fandom Transcripts Wiki (2021), a website where fans can upload transcripts of television shows for the public to access. Any transcripts not sourced from the Fandom Transcripts Wiki (2021) were found on other websites that provide free transcripts (e.g., scripts.com). Each transcript was edited by hand to correct any spelling and grammar errors that may have originated from non-professional transcription of the episodes.

Sesame Street

Data from *Sesame Street,* a high-rated educational children's television program that began airing in 1969, was selected to be part of our database of children's television media (Public Broadcasting Service, 2021a). *Sesame Street* was chosen due to its popularity among children in our target age group and accessibility of episode transcripts. Eleven different *Sesame Street* episodes, one *Sesame Street* direct-to-video special, and one *Sesame Street* feature film were selected as part of our sample. However, two episodes were excluded from the sample due to insufficient transcript length. Because the feature film *The Adventures of Elmo in Grouchland* was very long, it was split into three different sections.

Blue's Clues

A bulk of the data in our sample comes from popular children's show *Blue's Clues*. *Blue's Clues* is a long-running children's show franchise that began in 1996. *Blue's Clues* features

cut-out style animation that is reminiscent of children's books and depicts familiar scenes to children such as homelife (Forbes, 2006). Another similarity to children's books is the innate involvement of the viewers in each episodic narrative of the show (Forbes, 2006). This combination of familiar settings and elements was specifically designed by *Blue's Clues*' creators in order to foster a stable environment to facilitate learning in the everyday lives of children (Forbes, 2006). Due to its popularity and long runtime, transcripts for the series are freely available online. *Blue's Clues* transcripts are used heavily in our database as they are largely accessible and are thought to be representative of an average difficulty in language chosen for children's television media. The language utilized in *Blue's Clues* is modeled after everyday childhood life, so the transcripts were also assumed to be a representation of CDS, extending on prior research (Massaro, 2015a; 2017b; Grice, 1975) . In total, we included the transcripts of 143 different episodes of *Blue's Clues* from its six full seasons on air.

Daniel Tiger's Neighborhood

Daniel Tiger's Neighborhood, a children's television program based on the popular long-running show *Mister Rogers' Neighborhood*, was also included in our database of children's television media. *Daniel Tiger's Neighborhood* was chosen because it is aimed at teaching social and emotional skills to preschool children aged 2-4, and likely would be viewed by children in our target age group of 2-5 (Public Broadcasting Service, 2021b). In total, the transcript(s) for two episodes of *Daniel Tiger's Neighborhood* and *The Daniel Tiger Movie: Won't You Be Our Neighbor?* were included in the sample. Due to observed differences in length between these media, the *Daniel Tiger* transcripts were consolidated into one file and split into three equally-sized samples.

Butterbean's Café, Santiago of the Seas, Rainbow Rangers, and Bubble Guppies

The *Butterbean's Café* episode "Cricket Goes Camping!", *Santiago of the Seas* episode "The Legend of Captain Calavara", *Bubble Guppies* episode "The New Guppy!" and the *Rainbow Rangers* episode "Tree Hugger & Turtle in a Net" were also included as part of the sample. These episodes were chosen because their content is aimed at children in our target age range and the transcripts are freely available on the Nick Jr. YouTube channel.

Readability Analysis of Individual Children's Television Episodes

Each episode chosen to be part of our sample was individually evaluated and scored for text readability, though some were split into subsamples due to episode transcript length. In total, approximately 165 episodes of popular children's television shows and 3 children's direct-to-video and feature films were analyzed in our investigation. The readability scores for these individual episodes and movies are listed in their entirety in Appendix 2.

Results

Vocabulary

The current study builds on Massaro's (2017b) assessment of the linguistic and cognitive complexity of picture books, CDS, and ADS with measures of vocabulary and readability. To assess the vocabulary contained in databases of CDS, ADS, and picture books, words that did not occur in the 5,000 most frequently used spoken and written words from the Corpus of Contemporary American English (COCA, 1990- 2012) were counted. There were roughly three times as many rare word types in the picture book word corpus than in the CDS corpus and even

one and one-half times as many rare word types as ADS. We now analyze any vocabulary differences among TV Scripts, Picture Books, and CDS (Massaro, 2017b).

Our current databases differed in size and therefore it was necessary to equate the number of tokens across the three databases. Given that picture books had the smallest database (52,484 words), for the vocabulary analysis, we randomly sampled this number of tokens from the TV Scripts database (285,962 words) and the Parent utterances database (152,976 words).

These three samples were assessed against the 5,000 most frequently spoken and written words from the Corpus of Contemporary American English (COCA, 1990–2012). We also used normative word frequency measures from Brysbaert and New (2009), who compiled a new frequency measure of words on the basis of American subtitles (51 million words in total) from film and television. We used this subtlex database (SUBTLEX, 2021) to evaluate the normative frequency for the words in our three databases after eliminating all of the words that also occurred in the 5,000 most frequent words in the English language (as determined from COCA.

The token word samples from the TV Scripts, Picture Books, and Parent Speech databases were equal in size at 52,484 words. We first eliminated any words that did not occur in the subtlex word frequency database because these are most likely primarily non-words, unconventional spellings, and spellings with unknown meanings. Using this criterion, we eliminated 257, 482, and 425 words in the TV Scripts, Picture Books, and Parent Speech databases, respectively. In all 3 databases, the number of word types not occurring in the COCA database but also occurring in the subtlex database (SUBTLEX, 2021) was 2749, 4383, and 2893 for the TV Scripts, Picture Books, and Parent Speech databases, respectively. Thus, children are exposed to significantly more word types with picture books relative to the other two media.

Word frequencies for the three databases suggest that words within these databases occurred in the subtlex word frequency database between 1 and 1,057,301 times (SUBTLEX, 2021). If we count the number of words in each database that occurred between 1 and 200 times in the subtlex word frequency database, there are 1,413 words in the picture book database relative to just 650 words in the TV Scripts database and 579 words in the Parent Speech database. One might argue that this difference might reflect very infrequent words. To evaluate this possibility, we eliminated words with a frequency between 1 and 5 and words that occurred more often than 200 words in subtlex database (SUBTLEX, 2021). This reduced these values to 1276 rare words in the picture book database, 601 rare words in the TV Scripts database, and 546 rare words in the Parent Speech database. Therefore, children are gaining exposure to far more infrequent words in picture books that in the other two types of language input.

Although we have found more occurrences of rare words in picture books relative to CDS and TV Media, we haven't analyzed the relevance of these words for cognitive and linguistic development. One productive measure is to assess the value of these words for instruction in the classroom. Hiebert (2005) identified the words that accounted for 90% of total words in fourth-grade assessments of three states and the National Assessment of Educational Progress (NAEP, 2017). She found that 90% of the total words on all assessments were accounted for by the words 10 or more appearances per million words of text in the Zeno et al. (1995) Educator's Word Frequency Guide (EWFG), which was based on over 17 million words of texts that represented school content areas and grade levels from first through college. We found that picture books had about 2 or 3 times the number of these academic words than did CDS and TV

Media. The actual number of words meeting this criterion was 411 for picture books and only 142 for CDS and 168 for TV Media, respectively.

Words are also more valuable if they are found in different content areas. Dispersion is a measure that indexes to what extent a word occurs across subject areas. Dispersion value of 1 means a word occurs across all content areas and smaller values mean it occurs in fewer subject areas. The Vocabulary Assessment Study in Education (VASE, 2014) has an average dispersion level of .65. Thus, it seemed reasonable to assess the words in the three databases that had dispersion values larger than .65. Picture books provided about twice as many words with a dispersion value of .65 or greater than the other two media. The actual number of words meeting this dispersion criterion was 364 for picture books and only144 for CDS and 170 for TV Media, respectively.

We also chose to look at words with very high frequencies but did not occur in the COCA most frequent words based on the premise that these would not be particularly rare words. Applying the criterion of a frequency of 1000 or more in the subtlex word frequency wordlist, there were 526 words in the picture book database relative to 353 words in the TV Scripts database and 451 words in the Parent Speech database. These words cannot account for the large number of word types in the picture book database compared to the CDS and TV Media. Thus, children are simply being exposed to more word types in picture book reading but importantly these words tend to be important for future schooling and being present in a number of content areas.

Readability

Readability grade level was used to measure the quality of the language used in each of the three media: CDS. Picture Books, and TV Media. We use grade level to reflect the linguistic and cognitive complexity of speech in each database. Although, historically, readability measures were strictly reserved for text in books, more recent studies have shown the potential for utilizing readability measures for spoken language as well (Massaro, 2017b). To date, there exists a variety of readability formulas available that, when used in conjunction, putatively produce a generalizable estimate of grade levels for specific texts.

Various prominent readability formulas were applied to each sample and results were organized and averaged. The readability formulas that were used as part of our analysis are as follows: the Flesch Reading Ease Score, the Fog Scale (Gunning Fog), Flesch-Kincaid Grade Level, The Coleman-Liau Index, The SMOG Index, the Automated Readability Index, and the Linsear Write Formula.

To calculate the readability scores of each sample, each subsample was entered into the Readability Formulas website (https://readabilityformulas.com). This website used seven different readability formulas (listed in the Table 3) to generate seven readability scores and a "readability consensus" including overall grade level, reading level, and reader's age for each episode.

Insert Table 3 About Here

Readability scores from seven different readability formulas were calculated for each subsample in the CDS, picture book, and TV Media databases. Averages of these scores for the

three databases are depicted in Table 4. Picture books had the highest readability scores on average, with an average reading grade level of 3.5 compared to the average reading grade levels of CDS (reading grade level 2.0) and children's television media (reading grade level 2.5). Individual formula averages for each database reflected this advantage for picture books, with the picture book database generally having higher readability scores than CDS and children's television media for all formulas. The exception to this trend is scores produced by the Coleman-Liau Index, wherein picture books earned an average reading grade level of 5 and children's television media earned an average reading grade level of 5.3.

Insert Table 4 About Here

One trend consistent throughout all the formulas was children's picture books having higher readability scores on average than child directed speech. This robust linguistic advantage for picture books over CDS replicates previous findings by Massaro (2015a), which utilized the same sample of picture books but a different sample of child directed speech. As Massaro's (2015a) sample included only parents speaking to very young children (age about 8 months) and our sample included parents speaking to children aged 2-5, this finding demonstrates a consistency of results across different age groups.

Averages from Table 4 also show a small but consistent advantage for children's television media over child directed speech. This advantage appears across averages for all formulas, with children's television media consistently receiving higher readability scores than

child directed speech. This result suggests that children's television shows feature more advanced linguistic content than is found in child directed speech.

Appendix's 1-3 depict the readability averages for every sample within each database. (Expand more on these results).

Ultimately, our results indicate that samples of children's picture books are more linguistically and cognitively difficult on average than samples of children's television media and child directed speech (CDS). They also demonstrate that samples of children's television media are somewhat more linguistically and cognitively difficult on average than samples of child directed speech. However, overall averages from Table 4 point to all three modes of linguistic input for children being relatively reasonable in terms of grade level. Despite this, differences in average scores between sources of speech input imply a substantial advantage of children's picture books over CDS and TV Media. This suggests that while children's shows like *Sesame Street* and *Blue's Clues* may familiarize children with slightly more complex language than CDS, picture books are still a better source of complex language.

Retrospective

Opportunities for future research may include the creation of more recent and more expansive databases of children's picture books, CDS, and TV Media. While the databases in the present study were quite large, they can always be more representative of different media forms. In addition, some of the primary media sources are relatively dated. Future research can focus on more contemporary children's books and or TV Media, as it can be assumed that in the years since some of these books or shows have been created, there may be other popular media forms that could display even more varied readability scores. Indeed, the opportunities for seeking new

children's media is endless as new media is created specifically for childhood development and literacy everyday. Research can also focus on cultural themes and sources for such media, as multicultural books and shows are becoming increasingly popular for young families.

References

Bååth, R. A. (2010). ChildFreq. http://childfreq.sumsar.net/.

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, 977-

Daniel Tiger's Neighborhood (n.d.). Scripts.com. Retrieved July 14, 2021, from https://www.scripts.com/script/daniel_tiger's_neighborhood_1290.

DeLoache, J. S., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G. L., Strouse, G. A., & O'Doherty, K. (2010). Do Babies Learn From Baby Media? *Psychological Science*, *21*(11), 1570–1574. https://doi.org/10.1177/0956797610384145

Dingemanse, M., & Thompson, B. (2020). Playful iconicity: structural markedness underlies the relation between funniness and iconicity. *Language and Cognition*, *12*(1), 203–224. https://doi.org/10.1017/langcog.2019.49

Evans, M. A., & Saint-Aubin, J. (2013). Vocabulary acquisition without adult explanations in repeated shared book reading: An eye movement study. *Journal of Educational Psychology*, *105*(3), 596–608. https://doi.org/10.1037/a0032465

Forbes, J. (narrator) (27 July 2006). Behind the Clues: 10 Years of Blue (Part 1) (Short documentary). Nickelodeon. Retrieved 1 June 2021.

Grice, H. P. (1975). Logic and conversation. In A. P. Martinich (Ed.), Philosophy of language (pp.165-175). New York, NY: Oxford University Press.

Hiebert, E.H. (2005). In pursuit of an effective, efficient vocabulary curriculum for the elementary grades. In The Teaching and Learning of Vocabulary: Bringing Scientific

Research to Practice; Hiebert, E.H., Kamil, M., Eds.; LEA: Mahwah, NJ, USA, 2005; pp. 243–263.

Massaro, D. W. (2012). Acquiring Literacy Naturally: Behavioral science and technology could empower preschool children to learn to read naturally without instruction. *American Scientist, 100*, 324-333.

Massaro, D. W. (2015a). Two different communication genres and implications for vocabulary development and learning to read. *Journal of Literacy Research, 47*(4), 505-527. http://dx.doi.org.oca.ucsc.edu/10.1177/1086296X15627528

Massaro, D.W. (2015b). Speech Perception. In: James D. Wright (editor-in-chief), *International Encyclopedia of the Social & Behavioral Sciences*, 2nd edition, Vol 23. Oxford: Elsevier. pp. 235–242. ISBN: 9780080970868

Massaro, D. W. (2016). Multiple influences in vocabulary acquisition: Parental input dominates. Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016), pp. 878-882. ISSN 2308-457X. Available at www.isca-speech/archive/interspeech_2016/pdfs/0037.PDF

Massaro, D. W. (2017a). Modeling Multiple Influences on Vocabulary Acquisition: Context, Symbol, and Association Learning. *Unpublished paper.*

Massaro, D. W. (2017b). Reading aloud to children: Benefits and implications for acquiring literacy before schooling begins. *The American Journal of Psychology*, *130*(1), 63-72. http://dx.doi.org.oca.ucsc.edu/10.5406/amerjpsyc.130.1.0063

Massaro, D. W., & Perlman, M. (2017). Quantifying Iconicity's Contribution during Language Acquisition: Implications for Vocabulary Learning. Frontiers Communication, 09 March 2017 | https://doi.org/10.3389/fcomm.2017.00004

Massaro, D. W., & Rowe, B. (2015). Comprehension outscores production in language acquisition: Implications for Theories of Vocabulary Learning. *Journal of Child Language Acquisition and Development – JCLAD*, *3*(3), 121-152, 2015, September ISSN: 2148-1997

Montag, J.L., Jones, M.N., & Smith, L.B.(2015).The words children hear: Picture books and thes tatistics for language learning.Psychological Science. doi:10.1177/095679761559436

PBS Kids: Transcripts Wiki. Fandom. (n.d.). https://transcripts.fandom.com/wiki/Category:PBS_Kids.

Public Broadcasting Service. (2021a, August 24). *Daniel Tiger's Neighborhood*. PBS. https://www.pbs.org/parents/shows/daniel.

Public Broadcasting Service. (2021b, August 24). *Sesame Street*. PBS. https://www.pbs.org/parents/shows/sesame-street.

Public Broadcasting Service. (2021, August 24). *Daniel Tiger's Neighborhood*. PBS. https://www.pbs.org/parents/shows/daniel.

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2017). Iconicity in the speech of children and adults.. *Developmental science*.

Readability (accessed 2021). AUTOMATIC READABILITY CHECKER, a Free Readability Formula Consensus Calculator. Readability Formulas: AUTOMATIC READABILITY CHECKER. https://readabilityformulas.com/free-readability-formula-tests.php.

Shinskey, J. L. (2021). Lift-the-flap features in "first words" picture books impede word learning in 2-year-olds. *Journal of Educational Psychology*, *113*(4), 641–655. https://doi.org/10.1037/edu0000628

SUBTLEX (2021). https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus

TalkBank. (2003). CHILDES. Child Language Data Exchange System. https://childes.talkbank.org/.

The Daniel Tiger Movie: Won't You Be Our Neighbor? (2018) - Full Transcript. Subs like Script - all Movies and TV Shows Transcripts. (n.d.). https://subslikescript.com/movie/The_Daniel_Tiger_Movie_Wont_You_Be_Our_Neighbor-8847740.

U.S. Department of Education (2017). Academic Performance and Outcomes for English Learners: Performance on National Assessments and On-Time Graduation Rates; Author: Washington, DC, USA, 2017. Available online: https://www2.ed.gov/datastory/el-outcomes/index.html (accessed on September 5, 2021).

Vandewater, E. A., Bickham, D. S., & Lee, J. H. (2006). Time Well Spent? Relating Television Use to Children's Free-Time Activities. *Pediatrics*, *117*(2), e181–e191. https://doi.org/10.1542/peds.2005-0812

Vocabulary Innovations in Education Consortium. (2014). Vocabulary Assessment Study in Education; August 2014. Available online: Vocabulary Assessment Study in Education (VASE) (accessed on September 5, 2021).

Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D.,Valdez-Menchaca, M. C., Caulfield, M. (1988). Accelerating Language Development Through Picture Book Reading. *Developmental Psychology*, *24*(4), 552–559.

Zeno, S.M.; Ivens, S.H.; Millard, R.T.; Duvvuri, R. (1995). The Educator's Word Frequency Guide; Touchstone Applied Science Associates Inc.: Brewster, MA, USA.

Zimmerman, F. J., Christakis, D. A., & Meltzoff, A. N. (2007). Associations between Media Viewing and Language Development in Children Under Age 2 Years. *The Journal of Pediatrics*, *151*(4), 364–368. https://doi.org/10.1016/j.jpeds.2007.04.071

Table 1. Taxonomy Revealing the Potential Independence of Language Modality (Spoken vs. Written) and Formal (Nonconversational) Versus Informal (Conversational) Dialogue.

	Spoken Language Examples	Written Language Examples
	TED Talk	Non-Fiction book
Formal language	Lecture	Scholarly article
	MOOC*	Newspaper
	Face-to-face conversation	Texting
Informal language	TV dialogue	Instant messaging
	Fiction films	Light Fiction Writing

* Massive Open Online Course

Table 2. The number of word types that occurred in COCA, words not in COCA, and rare words defined as occurring between 6 and 200 tunes in subtlex.

	COCA	Not In COCA	Rare Words	
TVMedia	1285	1464	601	
Parent Utterances	1333	1560	546	
Picture Books	1634	2749	1276	

Table 3

Properties of the Transcripts and Utterances used in the children's speech database and the CDS database. We should make 2 tables. It does not make sense to analyze the proportions across CDS and children's speech. Also, if these proportions were used in the analyses they would be wrong. So we have to check on this.

e.

Subsample	Utterances	Transcripts	Proportion[a]
Male Age 2	8277	29	12.9%
Male Age 3	5476	24	8.5%
Male Age 4	4858	18	7.5%
Male Age 5	1726	8	2.7%
Male Total	20,337	79	31.6%
Female Age 2	4852	25	7.5%
Female Age 3	2827	15	4.4%
Female Age 4	7291	19	11.3%
Female Age 5	856	7	1.3%
Female Total	15,826	66	24.6%
Father	6,543	48	10.2%
Mother	21,698	101	33.7%
Parent Total	28,241	116	43.9%

Note. Number of transcripts and utterances for each sample and subsample.

[a] The percentage of utterances each condition and subsample has of the total number of sampled utterances (n=64,404), rounded to the nearest tenth.

Table 3

Seven Different Formulas to Assess Readability of Text

Name	Formula	Variables Used	Scoring Key
The Flesch Reading Ease Formula	RE = 206.835 - (1.015 x ASL) - (84.6 x ASW)	RE= Reading Ease ASL= Average Sentence Length ASW= Average Number of Syllables per Word	Outputs number from 1 to 100, with higher scores indicating higher reading ease. Scores 90-100: text can be understood by average fifth grader. Scores 60-70: text can be understood by average eighth/ninth grader. Scores 0-30: text can be understood by average college graduate.
The Flesch-Kincaid Grade Level Formula	FKRA = (0.39 x ASL) + (11.8 x ASW) - 15.59	FKRA= Flesch-Kincaid Reading Age ASL= Average Sentence Length ASW= Average Number of Syllables per Word	Outputs U.S. school grade level; grade level indicates that an average student in that grade level is capable of reading the text.
The Fog Scale	Grade Level = 0.4 (ASL + PHW)	ASL= Average Sentence Length PHW= Percentage of Hard Words (words with 3 or more syllables;	Score of 5: Readable Score of 10: Hard Score of 15: Difficult Score of 20: Very Difficult

		excludes proper nouns)	
The SMOG Index	SMOG grade = 3 + √ Polysyllable Count	Polysyllable Count= words with 3 or more syllables	Outputs U.S. school grade level; grade level indicates that an average student in that grade level is capable of reading the text.
The Coleman-Liau Index	CLI = 0.0588L - 0.296S - 15.8	L= Average Number of Letters per 100 Words S= Average Number of Sentences per 100 Words	Outputs U.S. school grade level; grade level indicates that an average student in that grade level is capable of reading the text.
Automated Readability Index	ARI = 4.71 (characters/words) + 0.5 (words/sentences) - 21.43	Word Difficulty= Number of Letters per Word (characters/words) Sentence Difficulty= Number of Words per Sentence (words/sentences)	Outputs U.S. school grade level; approximates grade level required to comprehend the text.
Linsear Write Formula	1. Calculate number of easy words; multiply by 1 2. Calculate number of difficult words; multiply by 3 3. Divide by the number of sentences 4. If answer is >20, divide by 2 for the final score	100 word sample Easy Words= 2 syllables or less Difficult Words=3 syllables or more Sentence Length= Number of Words in a Sentence	Outputs U.S. school grade level; approximates U.S. grade level of text sample.

	5. If the answer is < or = 20, subtract by 2 for the final score		

Note. Adapted from Brain Scott. (n.d.). AUTOMATIC READABILITY CHECKER, a Free Readability Formula Consensus Calculator. Accessed from https://readabilityformulas.com/

Table 4

Readability Scores and Averages for Samples of Children's Television Media, Children's Picture Books, and Child Directed Speech (CDS).

Sample	Flesch Reading Ease Score	Fog Scale	Flesch-Kincaid Grade Level	The Coleman-Liau Index	The SMOG Index	Automated Readability Index	Linsear Write Formula	Readability Consensus: Grade Level
Children's Television Media	96.78, very easy to read	3.3, easy to read	1.3, First Grade	5.3, Fifth Grade	3.26, Third Grade	1.13, 6–8 years old (First and Second Grade)	2.27, Second Grade	2.5, First, Second, and Third Graders
Children's Picture Books	91.7, very easy to read	4.8, easy to read	2.9, Third Grade	5, Fifth Grade	3.6, Fourth Grade	1.65, 6–8 years old (First and Second Grade)	4.16, Fourth Grade	3.5, Third, Fourth, and Fifth Graders
Child Directed Speech	99.38, very easy to read	3, easy to read	0.8, First Grade	4.5, Fifth Grade	2.67, Third Grade	-0.77, 3–5 years old (Preschool)	1.92, Second Grade	2.0 First and Second Graders
Average	95.95, very easy to read	3.67, easy to read	1.67, Second Grade	4.93, Fifth Grade	3.18, Third Grade	0.67, 6–8 years old (First and Second Grade)	2.78, Third Grade	2.67, Second, Third, and Fourth Graders

Note. Readability scores and averages for three main sources of speech input in early development.