Features of lexical richness in children's books: Comparisons with child-directed speech

Nicola Dawson
Yaling Hsiao
Alvin Wei Ming Tan
University of Oxford, UK

Nilanjana Banerji
Oxford University Press, UK

Kate Nation
University of Oxford, UK

Abstract: Access to children's books via shared reading may be a particularly rich source of linguistic input in the early years. To understand how exposure to book language supports children's learning, it is important to identify how book language differs to everyday conversation. We created a picture book corpus from 160 texts commonly read to children aged 0-5 years (around 320,000 words). We first quantified how the language of children's books differs from child-directed speech (compiled from 10 corpora in the CHILDES UK database, around 3.8 million words) on measures of lexical richness (diversity, density, sophistication), part of speech distributions, and structural properties. We also identified the words occurring in children's books that are most uniquely representative of book language. We found that children's book language is lexically denser, more lexically diverse, and comprises a larger proportion of rarer word types compared to child-directed speech. Nouns and adjectives are more common in book language whereas pronouns are more common in child-directed speech. Book words are more structurally complex in relation to both number of phonemes and morphological structure. They are also later acquired, more abstract, and more emotionally arousing than the words more common in child-directed speech. Written language provides unique linguistic input even in the pre-school years, well before children can read for themselves.

Keywords: lexical richness; book language; child-directed speech; language acquisition; literacy

Corresponding author: Nicola Dawson, Department of Experimental Psychology, University of Oxford, Anna Watts Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, United Kingdom. Email: nicola.dawson@psy.ox.ac.uk.

ORCID IDs: Nicola Dawson https://orcid.org/0000-0001-7167-6081; Yaling Hsiao https://orcid.org/0000-0003-3986-5178; Alvin Wei Ming Tan https://orcid.org/0000-0001-5551-7507; Kate Nation https://orcid.org/0000-0001-5048-6107

Citation: 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*, DOI: 10.34842/5we1-yk94

Introduction

Children learn from the language they hear (e.g., Cameron-Faulkner et al., 2003; Weisleder & Fernald, 2013). Evidence from longitudinal studies and computational modelling shows that children who experience greater amounts of sophisticated and diverse child-directed talk develop larger vocabularies and better reading skills, and are at an advantage in early school achievement (Chang & Monaghan, 2019; Hart & Risley, 1995; Hoff, 2003; Huttenlocher et al., 1991, 2010; Jones & Rowland, 2017; Pan et al., 2005; Rowe, 2008, 2012). Yet children's language experiences in the early years vary widely. These differences have been linked to caregiver language competence and socio-economic status (Hart & Risley, 1995; Hoff, 2003; Huttenlocher et al., 2010; Weisleder & Fernald, 2013), but language use may vary within, as well as between, home environments. Shared reading might be a particularly important source of language input, not least because it elicits more complex language and more words per minute from caregivers compared to other contexts, such as mealtimes and play (Demir-Lira et al., 2019; Weizman & Snow, 2001). In this paper we investigate in detail the language of children's books to specify the quantity and nature of lexical input they offer, relative to the language that children encounter via everyday speech.

Corpus analyses consistently demonstrate that written language departs from spoken language in several ways. These differences are well-documented in texts and speech aimed at adults. Overall, written language tends to be more syntactically complex and more lexically diverse than spoken language (Malvern et al., 2004; Roland et al., 2007), although patterns of language use may also reflect other factors, such as formality and genre (Biber, 1993). In part, linguistic differences across modality reflect the decontextualized nature of written language. As spoken language typically takes place in the 'here and now', communication is supported through gesture, facial expression and intonation. Spoken utterances that are incomplete or ambiguous may not pose a barrier to comprehension if meaning is apparent from the communication context. Speech may also be adapted in the moment to rectify breakdowns in communication (Clark, 2020; Healey, de Ruiter, et al., 2018; Healey, Mills, et al., 2018). In the absence of these nonverbal cues and bi-directional dynamics, written language depends more on choice of words and sentence structures to communicate information effectively (Snow, 2010).

Turning to children, books provide exposure to syntactic structures that occur rarely in speech. Montag (2019) showed that even in texts targeted at very young children (i.e. picture books), passive sentences and relative clauses occurred more frequently than in child-directed speech. Similar findings were reported by Cameron-Faulkner and Noble (2013), who found that canonical sentence structures (comprising subject-verb-[object]) and complex sentence constructions (containing two or more lexical verbs) were more frequent in children's books than child-directed speech, whereas questions were more common in speech than in books. Differences also emerge at the lexical level. Montag et al. (2015) calculated type-token ratio curves for a corpus of picture books and a corpus of child-directed speech, revealing that books contained more unique word types than speech at any given sample size. This pattern held true both at the corpus level, and in comparisons between individual books and conversations. Strikingly, even when compared to speech between two adults, children's picture books contain more unique rare word types (Massaro, 2015).

Together, these corpus comparisons suggest that children who frequently participate in shared

reading activities are regularly exposed to more advanced linguistic content than children who do not. These differences matter, given that language input is closely tied to language development and that regular access to books in the early years is not universal across children (Hart & Risley, 1995). Identifying and characterising common linguistic properties of children's books is an important starting point for understanding the impact of variation in access to books on children's language development. To this end, we introduce a new children's picture book corpus and identify critical properties of book language, focusing on its lexical content.

Lexical richness broadly refers to the quality of words in a language sample. It encompasses a number of measurable lexical properties, including lexical diversity, lexical density and lexical sophistication (Jarvis, 2013; Malvern et al., 2004; Read, 2000). Lexical diversity provides an indication of vocabulary breadth and is usually measured using type-token ratios (or type-token ratio curves; Montag et al., 2015, 2018). Lexical diversity tells us about the range of words in a text and has been widely adopted as a measure of language quality or proficiency (e.g., Malvern et al., 2004). Measures of lexical density capture the proportion of lexical items (usually defined as nouns, lexical verbs, adjectives and adverbs derived from adjectives) in a language sample relative to the total number of words (Ure, 1971). A higher proportion of lexical items in a language sample is indicative of denser information content compared to a sample with a higher proportion of function words (e.g., prepositions, conjunctions and pronouns). Lexical density is highly correlated with lexical diversity (Johansson, 2008), but conceptually, they measure distinct features. Hypothetically, it is possible for a text to have a high density of lexical items that are repeated frequently, or conversely, a text that uses a diverse range of vocabulary, but includes a high proportion of function words.

Like lexical density, measures of lexical sophistication shed light on the types of words contained within a language sample, and in particular, whether those words are skewed towards one end of the frequency distribution. One approach is to calculate the number of unique word types within a corpus after having accounted for the most frequent word types according to a general language corpus (Massaro, 2015). Adopting this method, Massaro reported that children's picture books contained around three times the number of rare word types of child-directed speech, and around one-and-a-half times the number observed in adult-adult speech. Alternatively, cumulative proportions of word tokens in a given corpus can be plotted against the rank frequency of those words in a general language corpus, providing additional information on the frequency distributions of the most common words across different corpora (Hayes, 1988; Hayes & Ahrens, 1988).

In summary, existing evidence indicates that children's books are more lexically diverse (Montag et al., 2015) and contain a higher proportion of rarer word types (Massaro, 2015) than child-directed speech. This indicates that the language of children's books is disproportionately skewed towards lexical items from the lower end of the frequency distribution. However, little is currently known about the properties of these words and lexical density has not been directly compared across these sources. This matters when we consider that children who are read to less frequently in the early years will gain less exposure to such words. Our aim here is to identify words that are relatively common in children's books, but which appear infrequently in child-directed speech, and to analyse their lexical properties. This will allow us to highlight the types of words that may be particularly impacted by variation in exposure to books in the early years.

Aims and Hypotheses

We created a new children's picture book corpus and selected samples of child-directed speech from the UK CHILDES corpora. This allowed us to fulfil three aims. Our first aim was to replicate Montag et al.'s (2015) analysis of lexical diversity in a new and larger set of children's picture books. Our second aim was to extend cross-modality comparisons to other measures of lexical richness (lexical density and lexical sophistication), along with part of speech distributions, word length, and morphological complexity. Our third aim was to identify the words most uniquely representative of children's books, and to examine how they differ from words more common in child-directed speech in relation to key psycholinguistic properties, namely age of acquisition (the age at which a word is learned; Kuperman et al., 2012), concreteness (the extent to which a word references a perceptible entity, or conversely, how abstract it is; Brysbaert et al., 2014), arousal (the intensity of emotion elicited by a word; Warriner et al., 2013), and valence (how pleasant a word is judged to be; Warriner et al., 2013). If the words most typical of books are more advanced and more abstract than words more common to child-directed speech, then children who regularly participate in shared reading activities will have more opportunity to encode the phonological forms and meanings of such words, and to experience them across diverse contexts. These experiences not only enhance oral vocabulary knowledge (Weizman & Snow, 2001), but also lay the foundations for reading development (Gough & Tunmer, 1986; Perfetti & Hart, 2002), even before children are able to read independently.

Following Montag et al. (2015), we predicted that our set of children's picture books would contain more diverse vocabulary than child-directed speech targeted at a similar age range. We also predicted that books would contain a higher proportion of content words, and more sophisticated vocabulary, relative to speech (Massaro, 2015, 2017). We further anticipated that differences would emerge in structural complexity and part of speech distributions. If the vocabulary of picture books is more sophisticated than that of child-directed speech, we would expect these words to be longer, and for books to contain a higher proportion of morphologically complex words. Given previous comparisons of written and spoken material in adult language samples, we expected differences to emerge in part of speech distributions across children's books and child-directed speech, in particular in the balance of nouns and pronouns (Biber et al., 1998; Hudson, 1994). Finally, we predicted that the words we identified as most representative of 'book language' would have a higher age of acquisition, and would be more abstract, more emotionally arousing, and evoke stronger positive and negative emotions than the words more typical of child-directed speech.

We present our findings in two parts. First, we describe our corpora and the methods used to compare lexical richness across book language and child-directed speech. We then introduce the keyword methodology used to identify words most and least representative of book language before comparing their psycholinguistic properties.

Method

Corpora and Corpus Processing

Picture Book Corpus

The picture book corpus comprised 160 children's fiction books with a total word count of 319,435. These books were purchased for the purposes of this research, and were selected to be representative of the type of reading material children encounter in shared reading contexts in the UK. To this end, we generated an initial list of titles with a target age range of 0-7 years from a combination of retailer bestseller lists and recommendations from literacy charities, book review sites, and teachers. The final list included the titles that were cited most frequently across these sources (see Appendix A for the final selection of book titles; the full corpus can be found at https://osf.io/zta29/). The vast majority of books in the corpus were picture books, but a small number of longer texts that might be read to young children were also included (e.g., The BFG). The content of these books was transcribed as plain text files by undergraduate psychology students. We included text that appeared in illustrations and appendages (for example, text in speech bubbles) in the transcription on the basis that caregivers would likely read these words aloud in addition to the main body of text.

The plain text files containing the transcribed picture books were converted to CHAT Transcription Format (.cha) files so that they could be processed using Computerised Language Analysis (CLAN) software (MacWhinney, 2000). The 'mor' function in CLAN was used to lemmatise and generate part-of-speech tagging for all words within the corpus. The output .cha files were then converted to XML and parsed using the XML package in R (R Developement Core Team, 2017), with the data outputted to .csv files, which were used in subsequent analyses.

Spoken Language Corpus

This was generated from 10 corpora from the English-UK section of the CHILDES database (MacWhinney, 2000). The sample comprised all suitable corpora from this collection, with the exception of those that focused on specific populations (e.g., children with language impairments). The final set of 10 corpora (see Appendix B; the full set of corpora are accessible via the link above) contained transcripts of interactions between 190 different children aged 6 weeks to 6 years and their caregivers, siblings, other family members and research investigators. Recordings took place across a variety of contexts, but typically involved structured and free play activities between children and their caregivers, as well as everyday routines such as mealtimes and bedtimes. Across all recordings, utterances produced by the child were filtered out, such that the final dataset contained only talk directed to the child for a total word count of 3,853,976. The CHILDES corpora were downloaded in CHAT format and had already been processed using CLAN. As above, these files were converted to XML and parsed using R, with data outputted to .csv files in the same format as the picture book corpus.

Procedure

(i) Corpus Comparisons

Lexical Diversity. Following Montag et al. (2015), we calculated type-token ratio curves to show the number of unique word types in each corpus at various token sample sizes. We took this approach because type-token ratios decrease as the number of tokens in a sample increases: the more words there are in a language sample, the more likely it is that words will be repeated (Montag et al., 2018). Because our spoken language corpus is considerably larger than our picture book corpus, it was not possible to compare the two corpora on a single measure of lexical diversity. We adopted Montag et al.'s (2015) method of calculating type-token ratios for multiple random samples from each corpus, ranging from 100 to 50,000 words in size, and increasing in increments of 100 words each time. One hundred simulations were generated at each sample size, each based on a new random sample, and type-token ratios were calculated as the mean type count across the 100 simulations divided by the sample size.

Lexical Density. Each lemma token was coded as 'lexical' or 'non-lexical'. Lexical lemmas were defined as nouns (excluding proper nouns and pronouns), adjectives, verbs (excluding modal verbs, such as 'do', 'will', 'can', 'must', 'shall', 'may', and auxiliary verbs 'be', 'have' and 'get') and adverbs derived from adjectives (e.g., 'fast' and 'happily'). All other tokens were coded as 'non-lexical'. We calculated lexical density by dividing the number of lexical items by the total number of lemmas in each individual text or conversation (Berman & Nir, 2010; Strömqvist et al., 2002).

Lexical Sophistication. Following Hayes (1988; also Hayes & Ahrens, 1988), we generated cumulative frequency curves showing the proportion of each corpus accounted for by the 1,000 most common words in English. We decided to use the SUBTLEX-UK database as our reference, which lists frequencies for around 160,000 words generated from subtitles of British television programmes (van Heuven et al., 2014). We chose this as our reference database for two reasons. Firstly, these frequencies have been shown to explain 4% more variance in word processing times than other large general language corpora (e.g., the British National Corpus; van Heuven et al., 2014). Secondly, we reasoned that television subtitles represent a hybrid between written and spoken language as they typically record scripted speech, and therefore this approach would not be biased towards one modality over the other.

Our analysis was based on the cleaned version of the SUBTLEX-UK frequency list, with digits and non-alphanumerical symbols removed. We further eliminated all proper nouns from the list, and then ranked the list by token frequency across all broadcasts and selected the top 1,000 words. We calculated the cumulative proportion of tokens in the picture book and spoken language corpora accounted for by the 1,000 most common words in the reference list. We noted some inconsistencies in the tokenised forms of words between the SUBTLEX list and our corpora processed by CLAN (for example, contracted forms such as *n't* in the word *wasn't* was listed as a token in the SUBTLEX list, but not in our corpora). This meant that a number of items in the 1,000 most common words returned a frequency of 0 or a very low frequency in the picture book and spoken corpora. Therefore, we checked all entries in the SUBTLEX list that occurred with 0 frequency in either corpus to ensure

that this was truly due to non-occurrence, and not inconsistency in tokenisation. In the case of inconsistency, we manually corrected the relevant entries in our corpora to align with the tokenised form in the SUBTLEX list. Finally, we plotted cumulative frequencies as a proportion of total corpus size against rank order of the 1,000 most common words.

Part of Speech. The automatic part of speech tags generated by CLAN were combined into broad lexical categories. For example, CLAN provides a unique tag for each different type of pronoun: these were reclassified for the purposes of our analysis as 'pronouns'. Our focus was on the major parts of speech, including nouns, lexical verbs, adjectives, adverbs, pronouns and determiners. All other tags, including modal and auxiliary verbs, proper nouns and communicators (e.g., 'ah') were coded as miscellaneous.

Structural Properties. We calculated word length in number of phonemes using the Carnegie Mellon Pronouncing Dictionary (Carnegie Mellon University, 2014) as the reference database. Data on number of phonemes were available for 84% words in the picture book corpus and 79% of the words in the spoken language corpus. We also recorded the morphological structure of the words in each text or conversation. We calculated the percentage of morphologically complex lemmas in each text or conversation (i.e. ignoring inflected word forms), and recorded whether complex words were derivations (e.g., *teacher*), compounds (e.g., *football*) or words that were formed through both compound and derivational processes (e.g., *footballer*). Our coding of morphological structure was based on information available in the MorphoLex (Sánchez-Gutiérrez et al., 2018) and MorphoQuantics (Laws & Ryder, 2014) databases. Lemmatised forms output by CLAN were checked for errors (e.g., stems that comprised only one segment of a compound – *foot* instead of *football*) and inconsistencies with lemmatised forms in our morphology reference databases (for example, we included inflectional suffixes in the lemmatised form of nouns derived from verbs – *the writing on the page* – or participle adjectives, such as *the painted bench*). Any identified errors or inconsistencies were manually corrected. Morphological information was available for 97% of the words in the picture book corpus and 95% of the words in the spoken language corpus.

(ii) Keyword Analysis

We followed the method outlined by Kilgarriff (2009; see also Kilgarriff, 2001) to identify the words most representative of the picture book corpus. We started by filtering out tokens tagged as proper nouns or letters, and we also removed tokens with missing part of speech information. We then mapped the remaining tokens to the list of corrected lemmas used in the analysis of morphology, with the exception that inflectional suffixes (-*ed* and -*ing*) were removed to align with lemmatised forms in the age of acquisition, concreteness and affective ratings (see below).

Taking the picture book corpus as the focus corpus, and the spoken language corpus as the reference corpus, we calculated a keyness score for each word that appeared in the former. The keyness score for a given word is the ratio of normalised frequency in the focus corpus to normalised frequency in the reference corpus. We used average reduced frequencies in place of raw frequencies to account for the dispersion of a word across the corpus. This is an adjusted frequency measure which is based on the distances between consecutive occurrences of a given word in a corpus (Hlaváčová, 2006; Savický & Hlaváčová, 2002). This approach addresses the issue of 'burstiness': words that occur with

high concentration within a small section of a corpus (e.g., within the same document), but sparsely elsewhere. Two words with the same raw frequency may differ on average reduced frequency if one is more evenly distributed across the corpus than the other. For a word that is completely evenly distributed, the average reduced frequency will be equivalent to the raw frequency.

A keyness score of 1 means that a word appears with equal frequency (per million) in each corpus, whereas a score greater than 1 indicates that the word occurs more frequently in the focus corpus than the reference corpus, and a score below 1 indicates that the word occurs less frequently in the focus corpus than the reference corpus. Given the problem of calculating ratios for words occurring in the focus corpus, but not at all in the reference corpus, we added a constant of 10 to all normalised frequencies before calculating keyness. We selected this value as the constant because it focuses the keyword analysis on the lower end of the frequency spectrum (Kilgarriff, 2009), which we considered to be important when identifying the words that children were unlikely to encounter in everyday conversation, but which they would experience through regular exposure to book language. We have included output from additional keyword analyses in Supplementary Materials (available on the OSF project page https://osf.io/zta29/) which focus on keywords in higher frequency ranges.

Once we had generated a keyness score for each item in the picture book corpus, we ranked them and selected the 500 words with highest keyness scores (i.e. the words most representative of the book language corpus; hereafter 'book+ words'), and the 500 words with the lowest keyness scores (the words least representative of books; hereafter 'book− words'). We chose to focus on 500 words from each end of the spectrum as this was approximately the largest sample for which all words in the book− set had a keyness score of less than 1, indicating that they occurred with greater relative frequency in the spoken language corpus compared to the picture book corpus. See Appendix C for a reduced list of the 50 book+ and 50 book− words with the most extreme keyness scores.

We then compared the two sets of words on a number of psycholinguistic properties to examine what characterises the words that children experience through book language, and how they differ to words more typical of child-directed speech.

Age of acquisition. We analysed the age at which our two sets of words are typically acquired using ratings from Kuperman et al. (2012). These norms are generated by asking adults to rate the age at which they think they learned a word, with lower ratings indicating that a word is acquired earlier in development.

Concreteness. This was based on ratings from adults (Brysbaert et al., 2014), where participants were asked to rate the extent to which a word refers to something perceptible (i.e. something that can be directly experienced via any of the five senses), or conversely, the extent to which a word's meaning is defined using other words. Ratings range from 1 for words that are highly abstract (e.g., *would*) to 5 for words that are highly concrete (e.g., *apple*).

Arousal. We examined emotional arousal using norms from Warriner et al. (2013). Participants in this study were asked to rate the intensity of emotion elicited by a given word, ranging from 1 for 'calm' (e.g., *librarian*) to 9 for 'excited' (e.g., *insanity*).

Valence. Our valence ratings were also taken from Warriner et al. (2013). These ratings indicate the extent to which a word evokes positive or negative emotions, and also range from 1-9 where 1 represents 'sad' (e.g., *murder*), and 9 'happy' (e.g., *sunshine*). Because our hypothesis relates to the extremity of valence ratings, rather than the direction of the effect, we transformed the mean valence rating for each word by centring it at the midpoint of the scale (i.e. 5, representing a neutral response), and calculating deviation from that point irrespective of direction. For example, a mean rating of 5 was allocated a score of 0, and mean ratings of 4 and 6 were each scored as 1.

Results

(i) Corpus Comparisons

Lexical Diversity

The mean number of word types at each sample size for the picture book and spoken language corpora are presented in Figure 1. The data show that, at any given sample size, the picture book corpus contains a greater number of unique word types than the spoken language corpus. Differences also emerge in the slopes of the lines. The picture book corpus shows a steeper type-token ratio curve compared to the spoken language corpus, indicating a greater increase in unique word types per unit increase in word tokens.

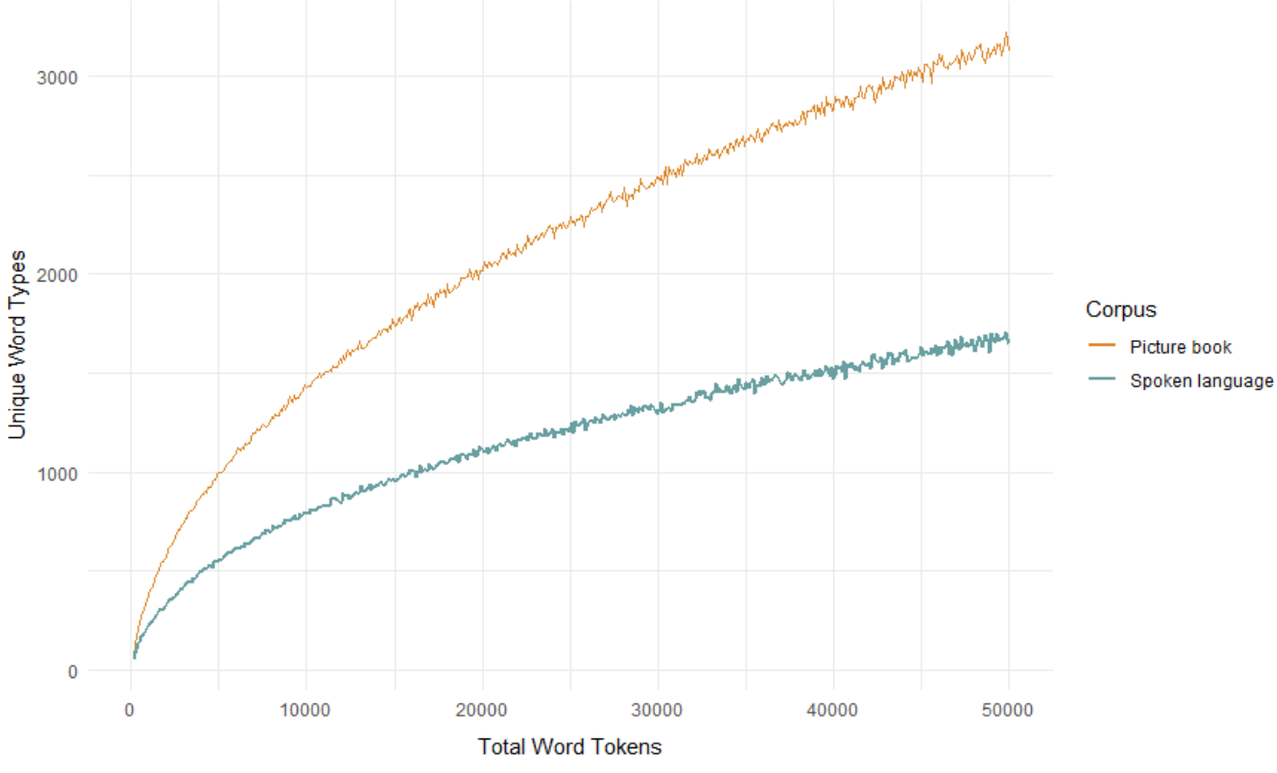

Figure 1. *Mean number of word types at different sized samples of word tokens randomly selected from the picture book and spoken language corpora*

Lexical Density

Figure 2 plots percentage lexical density for each individual text in the picture book corpus (n = 160), and each contiguous sample of child-directed speech in the spoken language corpus (n = 1616). The picture books contain a significantly higher percentage of content words (M = 43.77; SD = 7.00) compared to samples of child-directed speech (M = 28.56; SD = 2.65): $t(163.55) = 27.29$, $p < .0001$.

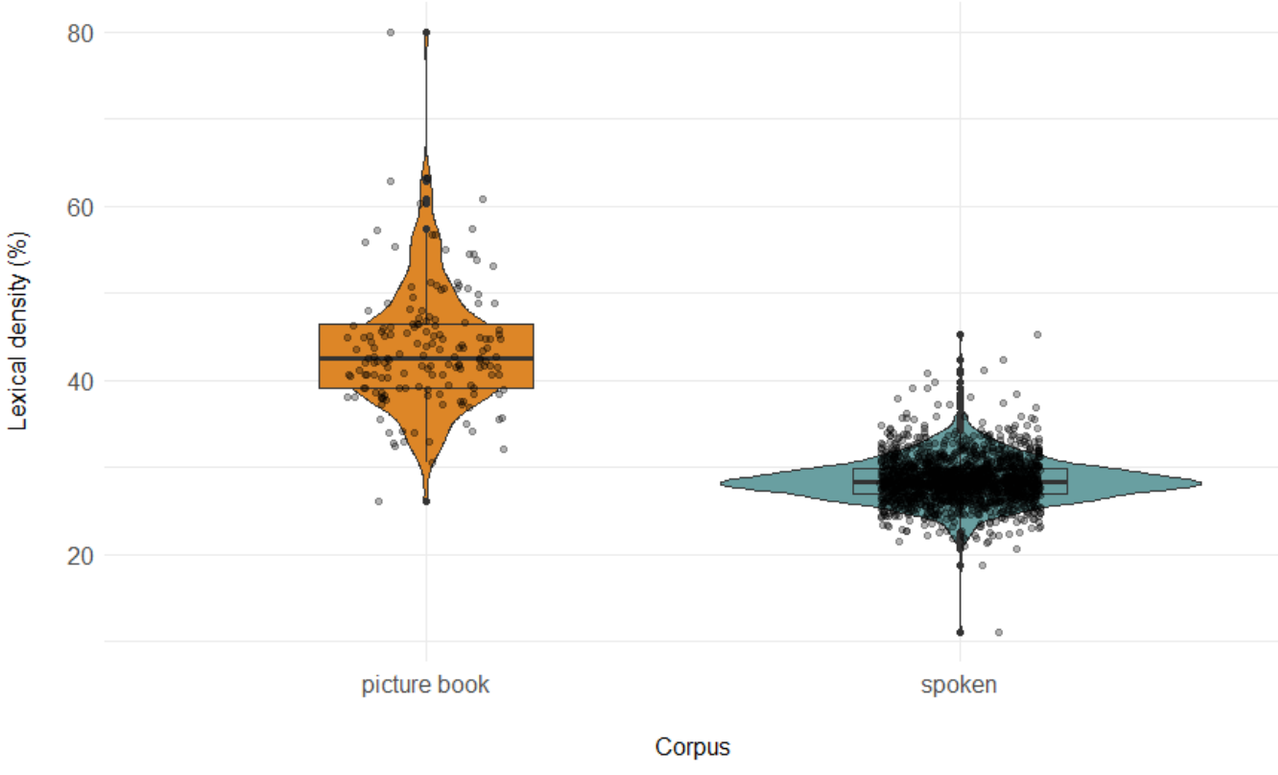

Figure 2. ***Percentage lexical density across picture book and spoken language corpora, plotted by individual document (picture book corpus) and conversation (spoken language corpus)***

We then examined whether lexical density varies by text genre. Specifically, we compared lexical density in texts written in a narrative style to those written in rhyme. It might be that rhyming texts would be more lexically dense than narrative texts, given the focus on imagery, rhythm and phonological properties of words. Texts adopting a partial rhyming structure were included in the 'rhyme' category, provided they were clearly written in verse. However, texts that were predominantly written in prose (e.g., a text comprising a collection of stories which included one story written in verse) were categorised as 'narrative'. Analysis revealed that percentage lexical density was indeed significantly greater in the rhyming texts (n = 62; M = 47.32; SD = 8.24) compared to the narrative texts (n = 98; M = 41.52; SD = 4.95): $t(89.03) = -4.99$; $p < .0001$ – see Figure 3). Inspection of the data

distributions indicated an outlier in the set of rhyming texts with a lexical density score of 80%. We reanalysed the data without this outlier, but this did not alter the outcome. Note that while lexical density was greater in the rhyming texts, narrative texts ($M = 41.52$; $SD = 4.95$) were still more dense than child-directed speech ($M = 28.56$; $SD = 2.65$).

Figure 3. ***Lexical density by text type***

Finally, we examined whether differences in lexical density across the book and spoken language corpora were driven by a proportionate increase across all lexical word classes, or a higher concentration of words from a particular word class. To do this, we calculated the frequency of nouns, verbs, adjectives and adverbs as a percentage of total lexical items in each corpus (Figure 4). If greater lexical density in the picture book corpus is equally distributed across word class, then there should be little difference across corpora in the frequency of each part of speech as a proportion of total lexical items. However, Figure 4 indicates a greater relative proportion of nouns and adjectives in the picture book corpus, and a lower proportion of verbs.

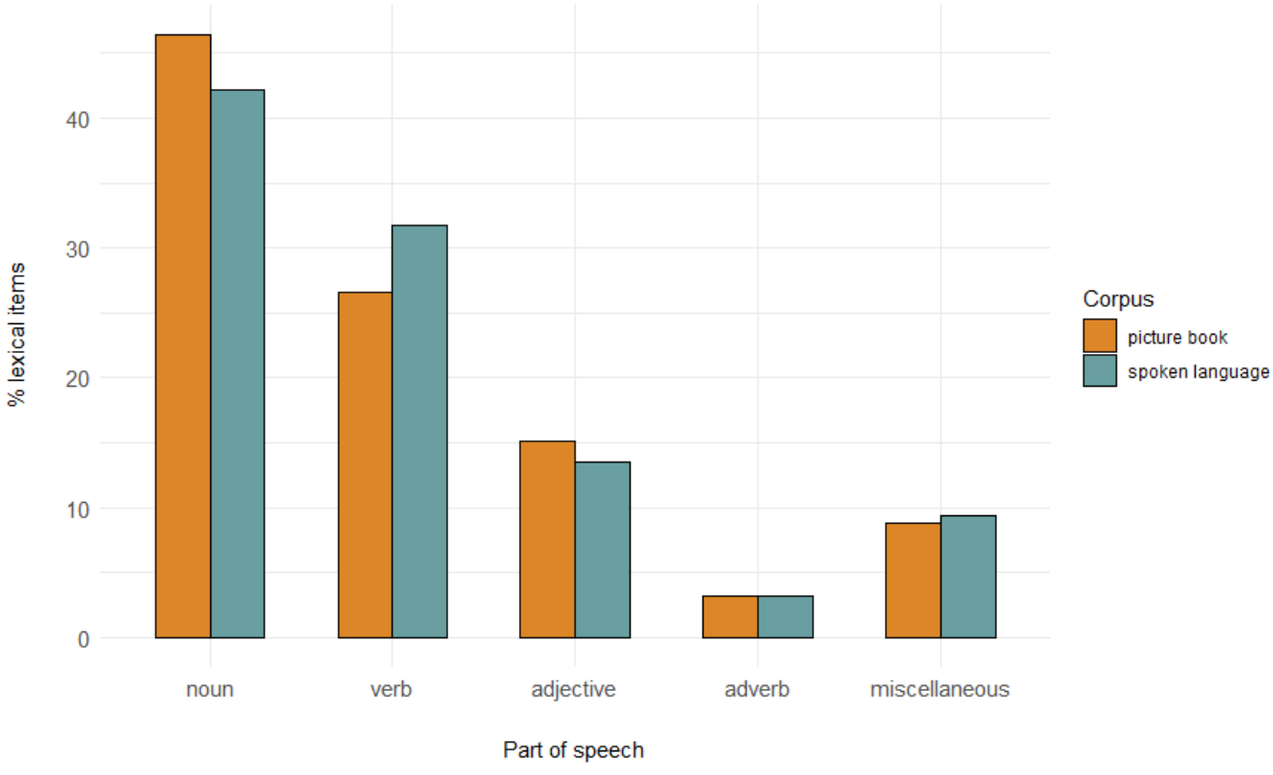

Figure 4. ***Frequency of part of speech tags as a percentage of total content words in the picture book and spoken language corpora***

Lexical Sophistication

Figure 5 plots the cumulative proportion of total tokens in each corpus accounted for by the 1,000 most common words in English (with SUBTLEX-UK television subtitles as the reference database), ranked in order of frequency on the log10 scale. The intercept at the left y-axis shows the proportion of each corpus accounted for the most common word according to SUBTLEX frequencies (*the*): 5% of the picture book corpus, and 3% of the spoken corpus. The point at which the curve intersects the right y-axis shows the proportion of each corpus accounted for by the 1,000 most common words: 72% of the picture book corpus, and 79% of the spoken corpus. The curves show that the words in picture books and child-directed speech are differently distributed along the frequency spectrum. A higher proportion of words in child-directed speech are among the most common words in the language overall, whereas picture books contain a higher proportion of words that fall outside this set. Therefore, access to picture books increases the likelihood that children will experience rarer word types that they would not otherwise encounter through conversation alone.

The curves also reveal an interesting pattern about the distributions of the most common words across the two modalities. As expected, the 1,000 most frequent words accounted for a larger proportion of total tokens in the spoken language corpus compared to the book corpus, yet the most

common words account for a higher proportion of words in the picture book corpus. Closer inspection of the top 10 words revealed that this effect was primarily driven by a higher proportion of articles (*the*, *a*) and conjunctions (*and*) in the book corpus, whereas the proportion of pronouns (*you*) and demonstratives (*that*) was greater in the spoken language corpus. We examine part of speech distributions in more detail next.

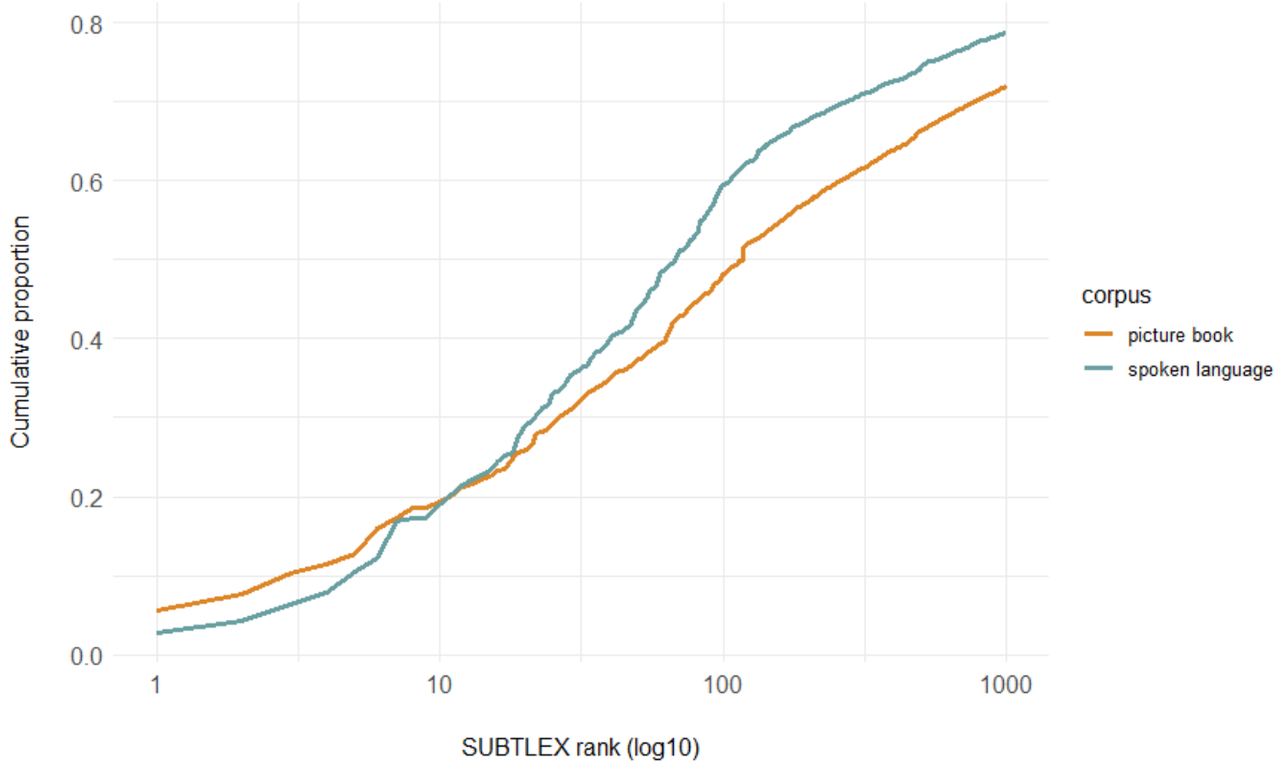

Figure 5. ***Cumulative proportions of total tokens plotted against rank of 1,000 most common words***

Part of Speech Distributions

Figure 6 shows frequency of occurrence (per million words) of each of the major lexical categories across the two corpora. Adjectives, conjunctions and coordinators, determiners, nouns, and prepositions all occurred with greater relative frequency in the picture book corpus compared to the spoken language corpus. Only pronouns were more frequent in spoken language, along with items classed as 'miscellaneous', which included proper nouns, auxiliary and modal verbs, and communicators (e.g., *ah*).

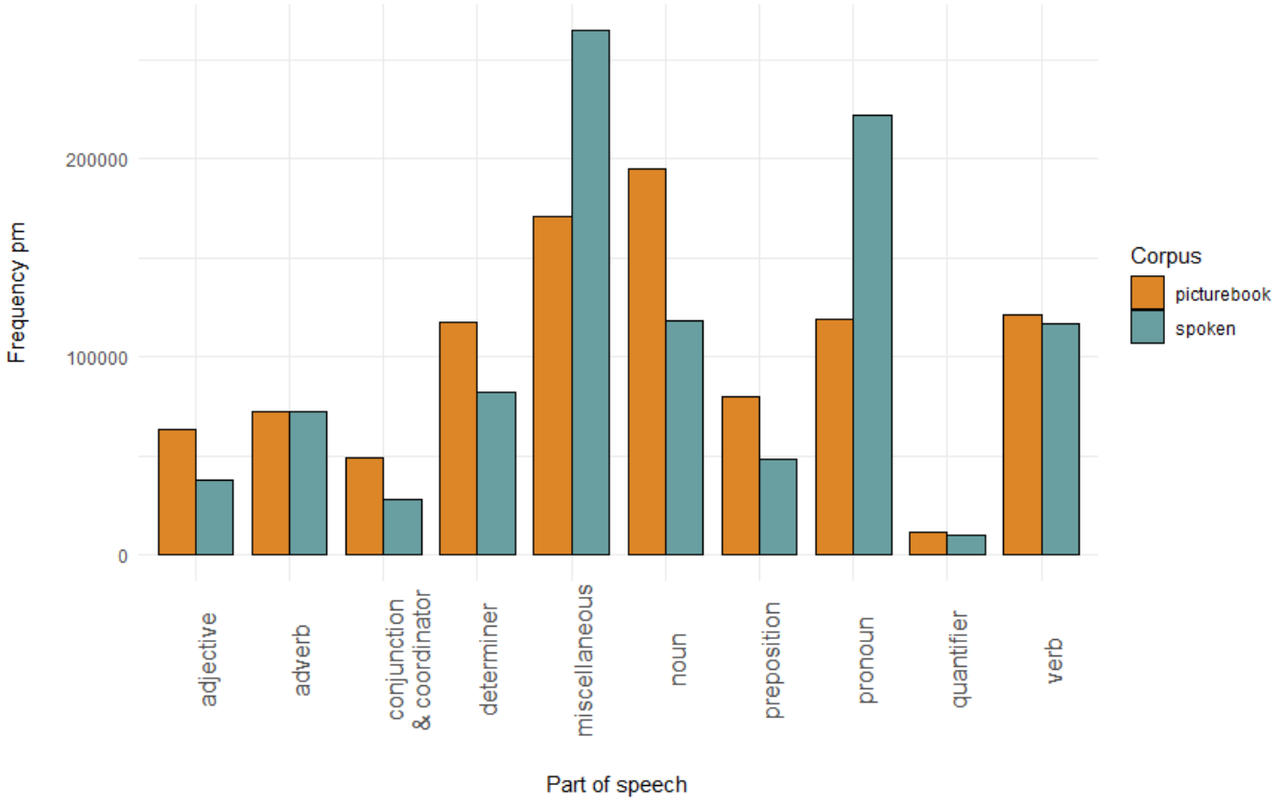

Figure 6. ***Part of speech distributions (frequency per million words) across picture book and spoken language corpora***

We conducted further analyses to examine the distributions of different types of pronoun and determiner across the two corpora. Figure 7 indicates that differences in pronoun frequency across picture books and child-directed speech are driven mostly by the large number of personal (*you*), demonstrative (*this*), and interrogative (*what*) pronouns in speech relative to books. While determiners are more frequent overall in books compared to speech, this is particularly the case for articles (*the*) and possessives (*her*), whereas demonstrative determiners (*these*), just as demonstrative pronouns, show the opposite trend.

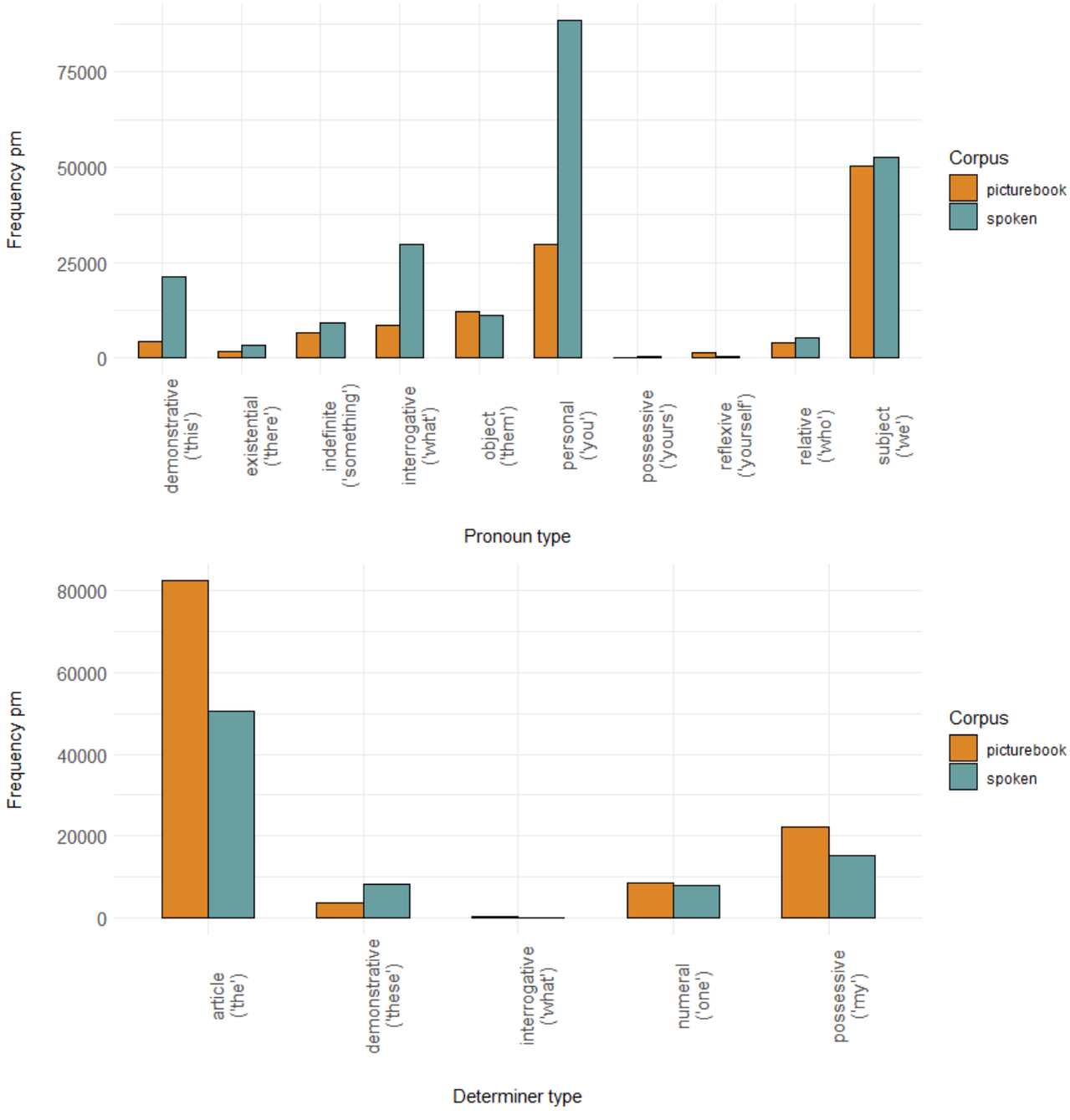

Figure 7. ***Pronoun (upper panel) and determiner (lower panel) distributions across picture book and spoken language corpora with examples from each category***

Word Length

Figure 8 shows phoneme count distributions across corpora. We set a maximum cut-off of 10 phonemes for the purposes of plotting the data, given the very small proportion of words that exceeded these values. The distributions indicate a higher proportion of longer words (four or more phonemes) in the picture book corpus, and a higher proportion of shorter words (three or fewer phonemes) in the spoken language corpus.

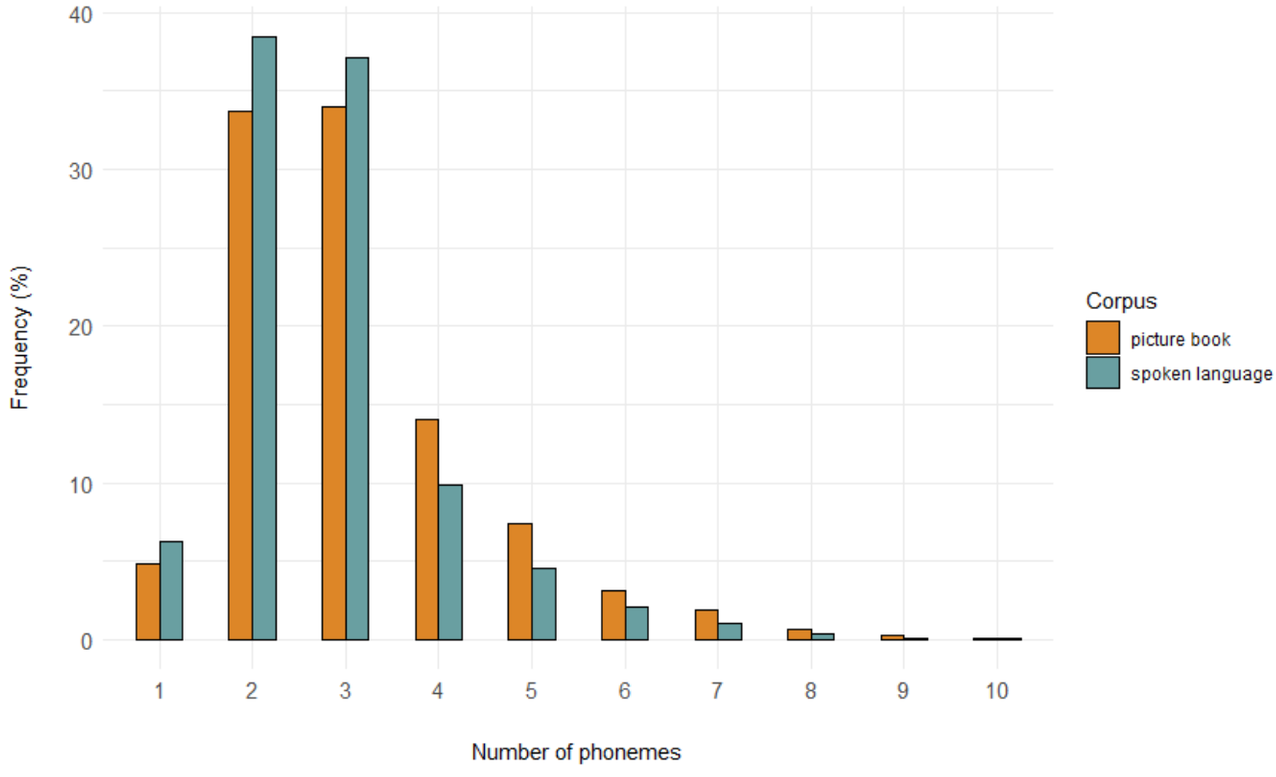

Figure 8. ***Phoneme count distributions across picture book and spoken language corpora***

Morphological Complexity

For each text or conversation, we calculated the percentage of morphologically complex lemma tokens (plotted in Figure 9). Plotting the full dataset indicated a number of outlier texts and conversations containing a high proportion of morphologically complex words (these were typically very short language samples). These were removed by excluding any individual text or conversation that exceeded three standard deviations from the mean for that corpus (corresponding to 0.63% of the texts in the picture book corpus and 0.43% of the conversations in the spoken language corpus). Removing these outliers did not alter the pattern of findings. Welch's Two Sample T-test confirmed that texts in the picture book corpus ($M = 6.61$; $SD = 3.19$) contained a significantly higher percentage of morphologically complex words than conversations in the spoken language corpus ($M = 4.31$;

SD = 1.09: *t*(161.68) = 9.03, *p* < .0001).

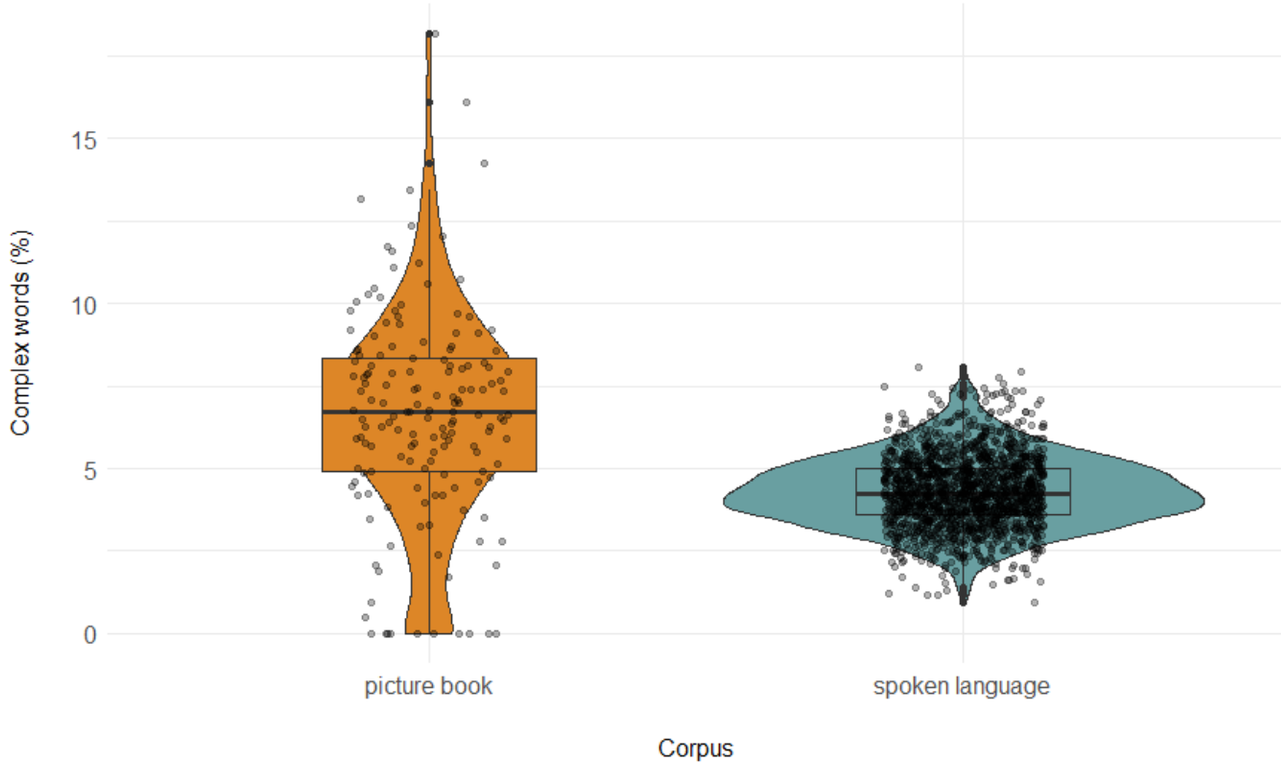

Figure 9. ***Percentage of words in each text (picture book corpus) or conversation (spoken language corpus) comprising two or more morphemes***

To further explore the composition of morphologically complex words across the picture book and spoken language corpora, we calculated the percentage of complex words accounted for by derivations and compounds. Figure 10 indicates that most morphologically complex words across the two corpora were derivations (e.g., *teacher*), followed by compounds (e.g., *football*), whereas derived compounds (e.g., *footballer*) were comparatively rare. The relative contribution of each word type to overall morphological complexity was very similar across the picture books and child-directed speech.

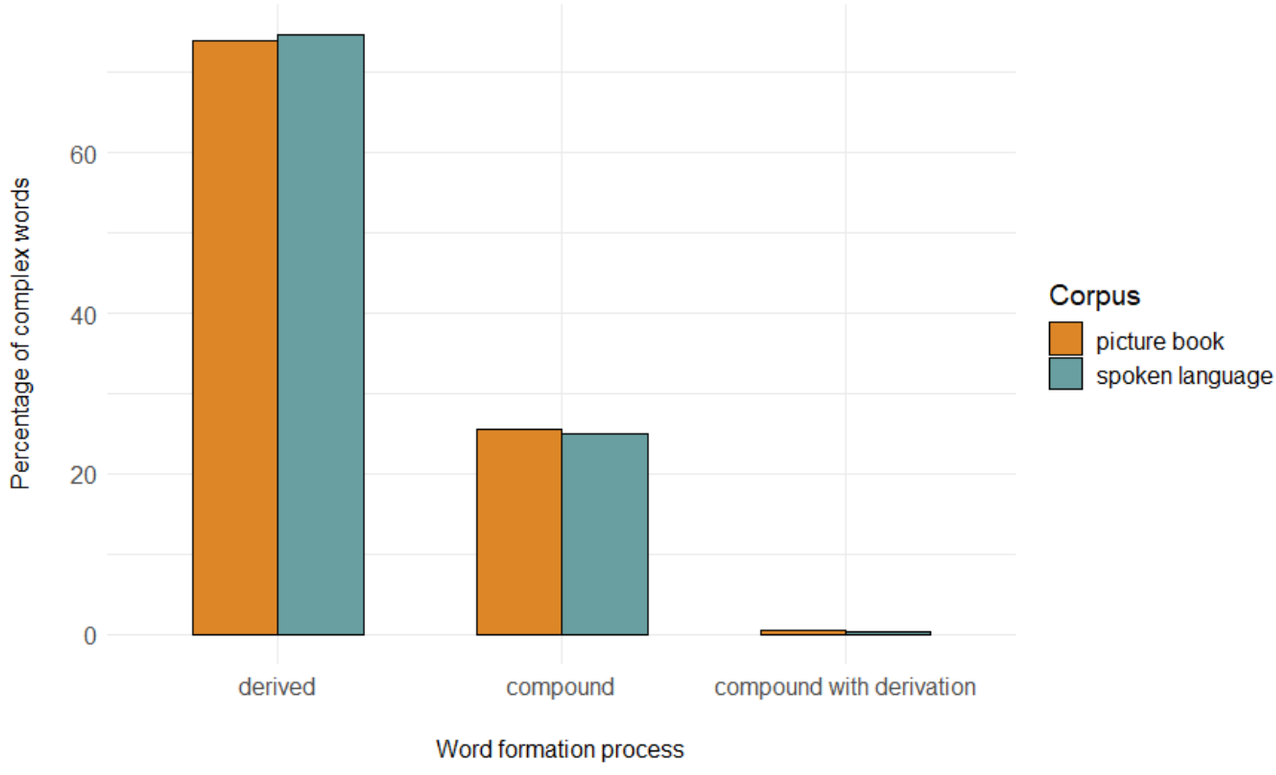

Figure 10. ***Percentage of total complex words in each corpus classed as derived, compound, and compounds with derivation***

(ii) Keyword Analysis

Age of Acquisition

Age of acquisition ratings were available for 462 of the 500 book+ words (*M* keyness score = 4.84, *SD* = 2.04), and 451 of the book− words (*M* keyness score = 0.65, *SD* = 0.21). Figure 11 shows distributions, box plots and data points for age of acquisition ratings for each set of words. Welch's Two Sample T-test indicated that the book+ words (*M* = 6.17; *SD* = 1.57) had a significantly higher mean age of acquisition rating than the book− words (*M* = 5.38; *SD* = 1.77): $t(892.63) = 7.11$, $p < .0001$).

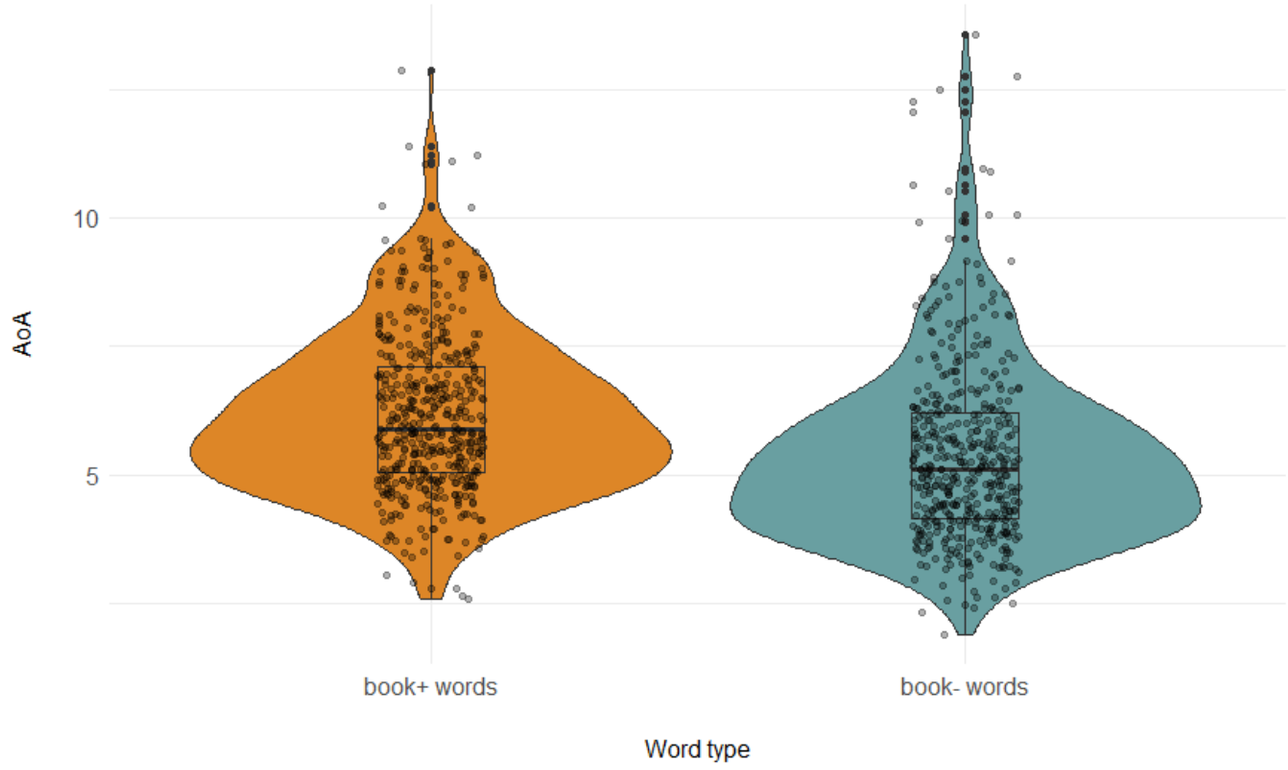

Figure 11. ***Age of acquisition ratings for the 500 words with the highest (book+) and lowest (book−) keyness scores***

Concreteness

Concreteness ratings were available for 491 of the book+ words (*M* keyness score = 4.82, *SD* = 2.00), and 469 of the book− words (*M* keyness score = 0.64, *SD* = 0.22). Figure 12 shows distributions, box plots and data points for concreteness ratings for each set of words. Welch's Two Sample T-test indicated that the book+ words (*M* = 3.27; *SD* = 0.98) are lower in concreteness than the book− words (*M* = 3.77; *SD* = 1.20): $t(901.59) = -6.99$, $p < .0001$).

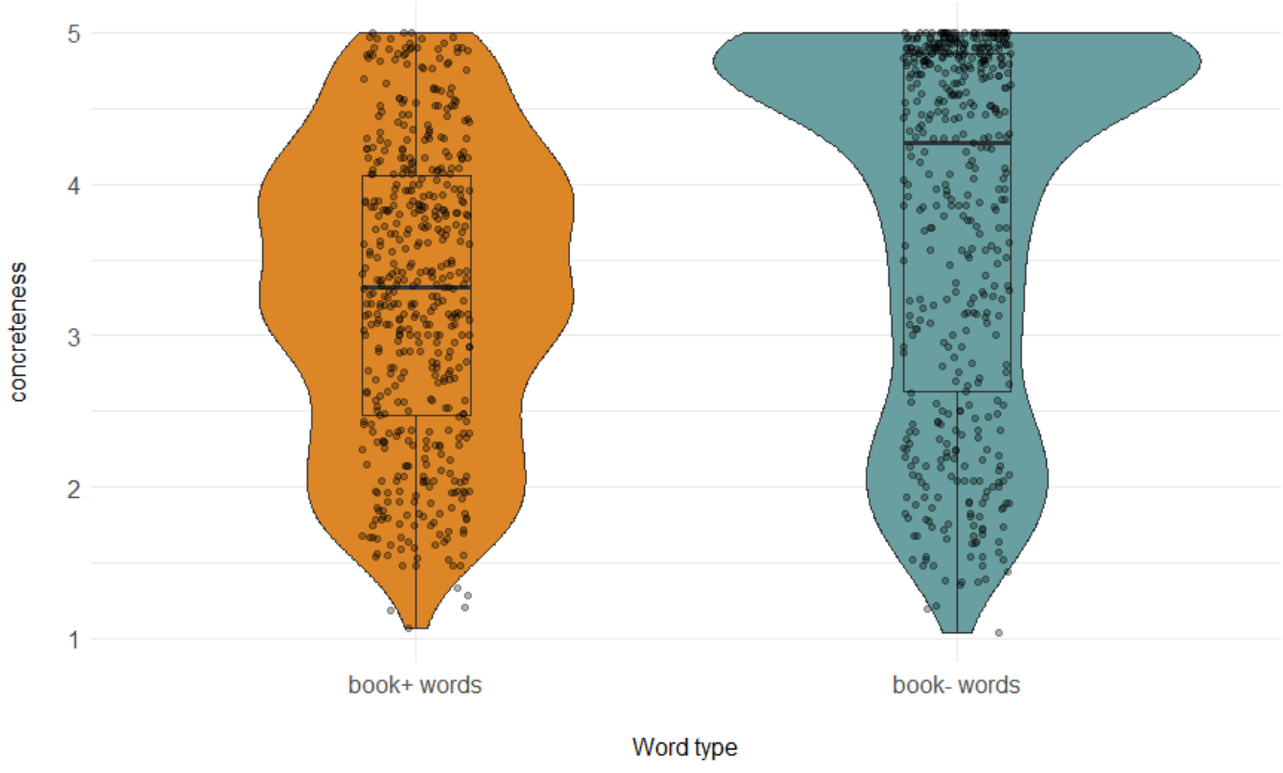

Figure 12. ***Concreteness ratings (max = 5) for the 500 words with the highest (book+) and lowest (book−) keyness scores***

Arousal

Arousal ratings were available for 389 of the book+ words (*M* keyness score = 4.82, *SD* = 2.06), and 365 of the book− words (*M* keyness score = 0.67, *SD* = 0.20). Figure 13 shows distributions, box plots and data points for arousal ratings for each set of words. Welch's Two Sample T-test indicated that the book+ words (*M* = 4.30; *SD* = 0.98) had a significantly higher arousal rating than the book− words (*M* = 3.98; *SD* = 0.83): $t(743.75) = 4.78$, $p < .0001$).

Figure 13. ***Arousal ratings (max = 9) for the 500 words with the highest (book+) and lowest (book−) keyness scores***

Valence

Valence ratings were available for the same words included in the analysis of arousal. Figure 14 shows distributions, box plots and data points for centred valence ratings for each set of words. Welch's Two Sample T-test indicated that there was no significant difference in the extremity of valence ratings between book+ words ($M = 1.21$; $SD = 0.82$) and book− words ($M = 1.15$; $SD = 0.70$): $t(745.79) = 1.04$, $p = 0.297$).

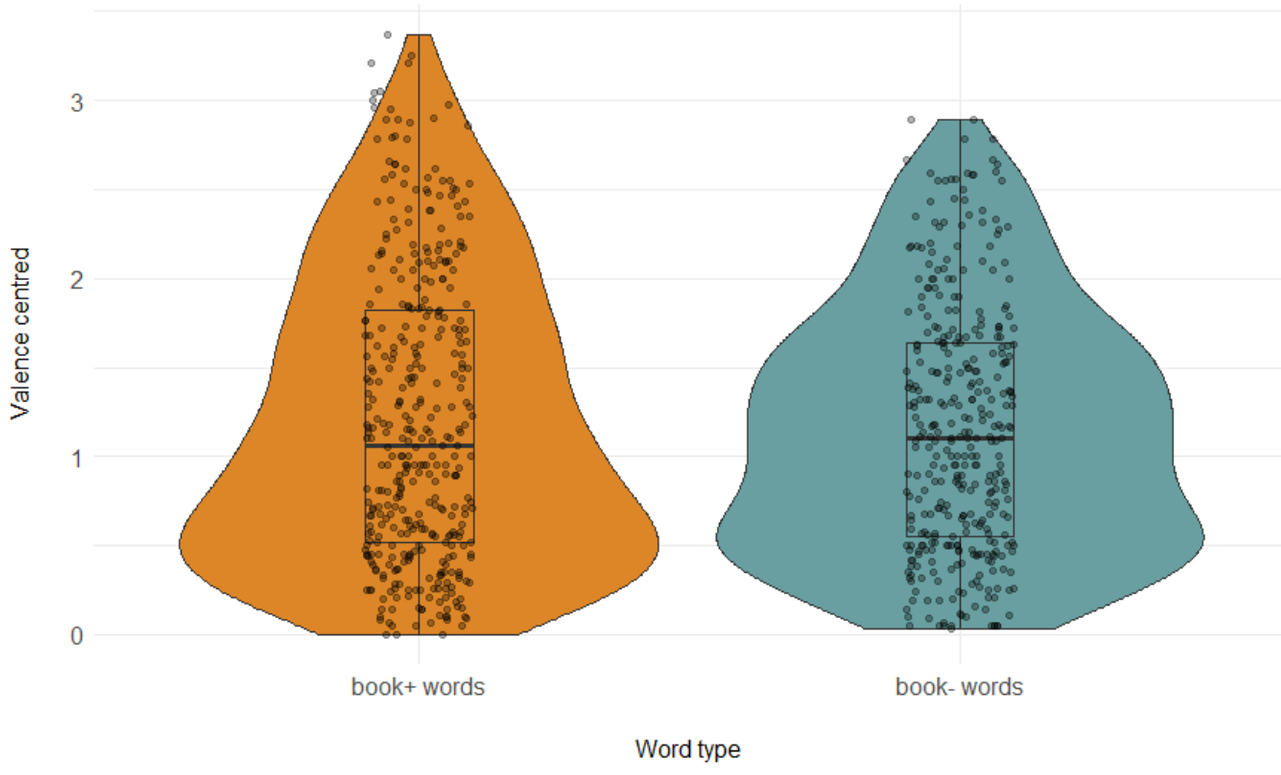

Figure 14. ***Centred (from the point of neutrality, see Method) valence ratings for the 500 words with the highest (book+) and lowest (book−) keyness scores***

Discussion

Our aim was to both replicate and build on previous work documenting differences in lexical richness across children's books and child-directed speech (Hayes, 1988; Massaro, 2015; Montag et al., 2015). In line with previous findings, we found that the words used in children's books are typically more diverse, more sophisticated, and lexically denser than those children hear via conversation. We extended these analyses by documenting the structural and lexical properties of these words. We found differences in part of speech distributions, with adjectives and nouns occurring more frequently in books, and pronouns more frequently in child-directed speech. The words in children's books were typically longer and were more likely to be morphologically complex, although the proportion of complex words that were formed through derivation or compounding was similar across the two corpora. Finally, we identified the words most representative of the books in our sample and found these had a higher age of acquisition, were more abstract, and rated higher in arousal than words more common to child-directed speech. We discuss each of these findings in turn and consider the implications for children's exposure to book language and language learning.

Following Montag et al. (2015), we compared lexical diversity in the picture book and spoken language corpora using type-token ratio curves. Our calculations were based on a different sample of

child-directed speech, and a new and larger corpus of children's books, yet our analyses clearly replicated their finding that picture books contain a greater number of unique word types than the spoken language corpus at any given sample size. Further, the curves representing type-token ratios showed a steeper trajectory for book language relative to spoken language. This suggests that increasing the amount of book language that children hear has a bigger impact on the number of unique words they are exposed to than an equivalent increase in child-directed speech. Diversity in the linguistic input is considered key to language learning (e.g., Johns et al., 2016). More specifically, some research suggests that lexical diversity in child-directed speech predicts children's vocabulary development over and above the quantity of language they hear (Hsu et al., 2017; Rowe, 2012), a finding backed by computational modelling separating the effects of quantity and diversity (Jones & Rowland, 2017). While caregiver talk may involve frequent repetitions of words and phrases in the context of regular routines, the words in books draw on a broader range of vocabulary sampled from a diverse set of topics. Not only do books provide children with access to these words, but they also provide a more contextually diverse environment for learning of individual words. Greater lexical diversity in the input means that a given word is more likely to co-occur with a broader range of other words, such that children have opportunities to develop semantic associations between them. Words that occur in more diverse contexts are acquired earlier in development, and show a processing advantage in older children and adults (Hills, 2013; Hills et al., 2010; Hsiao & Nation, 2018; Johns et al., 2016).

Our analysis of lexical diversity corroborates previous research showing that children encounter a broader range of vocabulary in books compared to an equivalent-sized input of child-directed speech. Turning to the types of words that children experience via books compared to conversation, our analyses of lexical density and lexical sophistication indicate that a higher proportion of the words in books are meaning-bearing words, and that they tend to occur less frequently in the language overall. This is important given that word frequencies are highly skewed, with only a small number of words occurring very frequently (predominantly function words) and the majority of words forming the long tail of the distribution (Piantadosi, 2014). Child-directed speech samples disproportionately from the higher end of this frequency spectrum. This is unsurprising because, unlike written language, speech is generated in the moment, and therefore word choice is biased towards those words in a speaker's lexicon that are most readily accessible (Navarrete et al., 2006). Similarly, because spoken communication incorporates extra-linguistic information, the variety, choice, and density of content words play a less crucial role in communicating meaning than they do in texts. This suggests that children's books are a particularly rich source of exposure to the types of words that children encounter rarely, if ever, in everyday conversation. While we focused on language directed primarily at pre-schoolers, children may have limited opportunity to access more advanced word types through speech alone, even once they reach school age: although caregivers draw on a more diverse vocabulary when speaking to older children, the types of words they choose come from the same part of the frequency distribution as the words used with younger children (Hayes & Ahrens, 1988). This evidence from older children reinforces book language as a critical source of lexical input.

Differences also emerged in part of speech distributions across the picture book and spoken language corpora. Our analysis revealed that among the major part of speech categories, nouns, adjectives, determiners, prepositions and conjunctions occur with greater relative frequency in books

compared to child-directed speech, whereas pronouns are almost twice as common in speech compared to books. The balance of nouns and pronouns in a language sample is typically a trade-off, given that they perform a similar grammatical function (Hudson, 1994). In most comparisons of written and spoken language, nouns are found to occur more frequently in texts than in speech, whereas the reverse is true for pronouns (Rayson et al., 2001). This pattern is particularly characteristic of informational or academic texts, where nominalisations are a common feature and occur more frequently than in fiction (Biber et al., 1998), but our findings indicate that the same is true even for fiction targeted at pre-school children. In books, explicit reference is important for comprehension: characters and objects do not exist in the immediate context and cannot be experienced directly. In child-directed speech, the focus of communication is more interpersonal and takes place within a shared context such that pronouns are often an adequate substitute for nouns. The breakdown of pronoun types indicates that differences in frequency were particularly stark in relation to demonstrative (e.g., *that's the wrong one*), interrogative (*what did I say?)* and personal (*you'll get stuck*) pronouns, all of which reflect a more involved and interactive style and reference entities within the immediate physical environment.

Adjectives were also more characteristic of books than of child-directed speech. Again, this finding aligns with comparisons of written and spoken language more broadly: given that adjectives modify nouns, a greater proportion of nouns in a text is likely to be accompanied by a similar rise in adjectives (Biber, 1988; Mair et al., 2002; Rayson et al., 2001). Nevertheless, in relation to children's learning, acquisition of adjectives plays a key role in the development of a sophisticated lexicon. Adjectives form the basis of descriptions (e.g., *the fluffy cat*) and contrastive relations (e.g., *big truck vs. little truck*), and provide linguistic labels for sensory perceptions, values, and emotions (e.g., *she is cold; he is good; I feel happy*). The meanings of adjectives also tend to vary according to context. For example, *a big rat* differs in size to *a big building* – such terms are relative rather than absolute (Davies et al., 2020). Therefore, experiencing an adjective in combination with a more diverse set of nouns may facilitate a more robust and flexible representation of that word (Blackwell, 2005). This contextual dependency also suggests that children need some basic knowledge of the nouns being modified by a given adjective before they can develop mastery of the adjective itself. Unsurprisingly, children learn adjectives at a slower rate than they do other open word classes, particularly nouns (Caselli et al., 1995; Gasser & Smith, 1998; Sandhofer & Smith, 2007). Storybooks may be a particularly rich source of input for acquisition of adjectives, given that they occur more frequently than in speech, and they also provide more varied contexts through which semantic representations of adjectives can be accumulated and refined.

Our keyword analysis revealed the words that were most unique to the books in our corpus, and a second set of words that occurred in the books, but were relatively more frequent in child-directed speech. We found that the words most representative of children's books are typically acquired later in development according to age of acquisition norms, and are more abstract and more emotionally arousing than the words more common in child-directed speech. However, we found no difference between the two sets of words in relation to whether the emotions they evoked were strongly positive or negative. These findings corroborate our analysis of lexical sophistication, showing that the words in books are more advanced not only in terms of their frequency of occurrence in English overall, but also in relation to the stage of development that children usually acquire them. This has implications for children's language learning. Words that are acquired earlier in development tend

to be well-connected to other words in the lexicon, whereas later words have fewer connections (Hills et al., 2009; Steyvers & Tenenbaum, 2005). According to one theory of how children expand their semantic network, the order in which children acquire new words reflects the connectivity of those words to other words in the learning environment (Hills et al., 2009). The words children hear in child-directed speech have a lower age of acquisition on average, and are more likely to be well-integrated in children's semantic networks. Access to books, on the other hand, provides an environment in which children can build semantic associations and develop connections between words that they may not otherwise encounter until later in development.

These words will also typically be more abstract. Concreteness is an important predictor of lexical processing in adults, with words higher in concreteness showing an advantage over abstract words (e.g., Binder et al., 2005), and abstract words tend to be acquired later in development (Ponari et al., 2018). One explanation is that concrete words (e.g., *apple*) refer to concepts that encode direct sensory experiences, and these imaginal representations are activated alongside verbal information during processing and retrieval. By contrast, abstract words (e.g., *validity*) rely more heavily on semantic information encoded linguistically, and the absence of support from perceptual memory means that these words are processed less efficiently (Paivio, 1971, 2013). The concreteness effect has also been accounted for by differences in contextual availability: abstract words are more challenging because they have weaker connections to associated contextual information, which makes it more difficult for an individual to activate that information when the word is encountered in isolation (Schwanenflugel, 1991). Underpinning both accounts is the idea that linguistic experience is key to the acquisition and processing of abstract words. Our analyses suggest that books provide more concentrated access to the types of words that are not supported by direct sensory experience, along with the linguistic and contextual information needed to support learning and consolidation. Acquisition of these words may be supported too by their affective properties. We found that the words in picture books were more emotionally arousing than the words in child-directed speech, although they did not differ on strength of valence ratings. Some theories of embodied semantics propose that emotion may play an important role in the acquisition and processing of abstract words in particular, functioning as an alternate source of experiential information in the absence of sensorimotor input (Kousta et al., 2011; Ponari et al., 2018; Vigliocco et al., 2014, 2018). However, a recent cross-linguistic study based on data from the MacArthur-Bates Communicative Development Inventory (Fenson et al., 2007) found limited evidence that arousal and valence predicted children's comprehension and production of early-acquired words (Braginsky et al., 2019).

Our comparisons of children's picture books and child-directed speech provide clear evidence that books are lexically richer overall, and have a different composition in relation to grammatical class and structural complexity compared to speech. Furthermore, the words children are least likely to encounter via conversation alone are more advanced, more abstract, and more emotionally arousing. Many of the features of 'book language' we have identified are true of written vs. spoken language comparisons more broadly (e.g., Biber, 1988), but it is nevertheless important to document the ways in which these sources of language input differ in relation to children's experiences. Doing so not only highlights the specific lexical structures and properties that may vary across language learning environments, but also reveals that even books designed to be accessible to the youngest children still provide a rich lexical input that is quite different to everyday speech.

In many ways, narrative fiction, particularly for young children, is more akin to oral language than other written genres (e.g., academic texts, newspapers), meaning that our findings are likely to be conservative estimates of the differences between book language and speech. However, it is also important to recognise that the corpus of child-directed speech we used here was predominantly sampled from interactions taking place within home settings, and that this may have limited the range and richness of vocabulary that caregivers used with their children. For example, experiences outside the home (a visit to the zoo, a trip to the beach) may provide greater opportunity for novelty and variety in lexical use, and for talk beyond the 'here and now'. More broadly, while corpus data provides valuable insights into the language structures children have opportunities to experience via books, it cannot speak to the effects of exposure on learning in individuals. Frequency counts alone do not capture the rich, interactive contexts in which language learning takes place (Roy et al., 2015), and nor do they accommodate the wider benefits of shared reading experiences, such as extra-text talk, scaffolding and emotional bonding.

While less lexically rich than book language, child-directed speech nevertheless plays an important part in children's language development. Certain properties of child-directed speech, such as exaggerated intonation patterns and grammatical simplification, have been hypothesised to support early language acquisition (Soderstrom, 2007). Given that the words in books are more advanced, the impact of variation in exposure to book language may relate more closely to the skills that underpin children's emerging literacy. The words that children encounter in picture books are by definition more characteristic of the literary domain. Importantly, experience is key: exposure to picture books via shared reading allows children to start encoding the phonological forms and meanings of more advanced words across different contexts from an early age. Over time, this experience will shape language development and provide a strong foundation to literacy (e.g., Gough & Tunmer, 1986; Perfetti & Hart, 2002). While there are many potential benefits of shared reading for children's development, our findings suggest that one of the key contributions may stem from the language of the books themselves, and specifically the rich and diverse lexical input they offer.

References

Berman, R. A., & Nir, B. (2010). The lexicon in writing–speech-differentiation: Developmental perspectives. *Written Language and Literacy*, *13*(2), 183–205. https://doi.org/10.1075/wll.13.2.01ber

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

Biber, D. (1993). Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics, 19*, 219–241.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17*(6), 905–917. https://doi.org/10.1162/0898929054021102

Blackwell, A. A. (2005). Acquiring the English adjective lexicon: Relationships with input properties and adjectival semantic typology. *Journal of Child Language, 32*(3), 535–562. https://doi.org/10.1017/S0305000905006938

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and Variability in Children's Word Learning Across Languages. *Open Mind, 3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand English word lemmas. *Behavior Research Methods, 46*(3), 904–911. https://biblio.ugent.be/publication/5774089/file/5774125.pdf

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science, 27*(6), 843–873. https://doi.org/10.1016/j.cogsci.2003.06.001

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

Carnegie Mellon University. (2014). *The CMU Pronouncing Dictionary*. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10*(2), 159–199. https://doi.org/10.1016/0885-2014(95)90008-X

Chang, Y. N., & Monaghan, P. (2019). Quantity and Diversity of Preliteracy Language Exposure Both Affect Literacy Development: Evidence from a Computational Model of Reading. *Scientific Studies of Reading, 23*(3), 235–253. https://doi.org/10.1080/10888438.2018.1529177

Clark, E. V. (2020). Conversational Repair and the Acquisition of Language. *Discourse Processes, 57*(5–6), 441–459. https://doi.org/10.1080/0163853X.2020.1719795

Davies, C., Lingwood, J., & Arunachalam, S. (2020). Adjective forms and functions in British English child-directed speech. *Journal of Child Language, 47,* 159–185. https://doi.org/10.1017/S0305000919000242

Demir-Lira, E., 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*, 1–16. https://doi.org/10.1111/desc.12764

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *The MacArthur-Bates Communicative Development Inventories User's Guide and Technical Manual (2nd Edition)*. Brookes Publishing.

Gasser, M., & Smith, L. B. (1998). Learning Nouns and Adjectives: A Connectionist Account. *Language and Cognitive Processes, 13*(2–3), 269–306. https://doi.org/10.1080/016909698386537

Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. *Remedial and Special Education, 7,* 6–10.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hayes, D. P. (1988). Speaking and writing: Distinct patterns of word choice. *Journal of Memory and Language, 27*, 572–585. https://doi.org/10.1016/0749-596X(88)90027-7

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

Healey, P. G. T., de Ruiter, J. P., & Mills, G. J. (2018). Editors' Introduction: Miscommunication. *Topics in Cognitive Science, 10*(2), 264–278. https://doi.org/10.1111/tops.12340

Healey, P. G. T., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science, 10*(2), 367–388. https://doi.org/10.1111/tops.12336

Hills, T. (2013). The company that words keep: Comparing the statistical structure of child- Versus adult-Directed language. *Journal of Child Language, 40*(3), 586–604. https://doi.org/10.1017/S0305000912000165

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

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science, 20*(6), 729–739. https://doi.org/10.1111/j.1467-9280.2009.02365.x

Hlaváčová, J. (2006). New approach to frequency dictionaries - Czech example. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, 373–378.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 17*(5), 1368–1378. https://doi.org/10.1111/j.1467-8721.2008.00596.x

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language, 103*(August), 114–126. https://doi.org/10.1016/j.jml.2018.08.005

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

Hudson, R. (1994). About 37% of Word-Tokens are Nouns. *Language, 70*(2), 331–339.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early Vocabulary Growth: Relation to Language Input and Gender. *Developmental Psychology, 27*(2), 236–248. https://doi.org/10.1037/0012-1649.27.2.236

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61*(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning, 63*(SUPPL. 1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics, 53*, 61–79. https://doi.org/10.7820/vli.v01.1.koizumi

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin and Review, 23*(4), 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology, 98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics, 6*(1), 97–133.

Kilgarriff, A. (2009). Simple Maths for Keywords. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of Corpus Linguistics Conference CL2009*.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The Representation of Abstract Words: Why Emotion Matters. *Journal of Experimental Psychology: General, 140*(1), 14–34. https://doi.org/10.1037/a0021446

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990. https://doi.org/10.3758/s13428-012-0210-4

Laws, J., & Ryder, C. (2014). *MorphoQuantics*. http://morphoquantics.co.uk/

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition*. Lawrence Erlbaum Associates.

Mair, C., Hundt, M., Leech, G. N., & Smith, N. (2002). Short Term Diachronic Shifts in Part-of-Speech Frequencies: A Comparison of the Tagged LOB and F-LOB Corpora. *International Journal of Corpus Linguistics, 7*(2), 245–264. https://doi.org/10.2307/3722566

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan.

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

Massaro, D. W. (2017). Reading aloud to children: Benefits and implications for acquiring literacy before schooling begins. *The American Journal of Psychology, 130*(1), 63–72.

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

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and Diversity: Simulating Early Word Learning Environments. *Cognitive Science, 42*, 375–412. https://doi.org/10.1111/cogs.12592

Navarrete, E., Basagni, B., Alario, F. X., & Costa, A. (2006). Does word frequency affect lexical selection in speech production? *Quarterly Journal of Experimental Psychology, 59*(10), 1681–1690. https://doi.org/10.1080/17470210600750558

Paivio, A. (1971). *Imagery and Verbal Processes*. Holt, Rinehart and Winston.

Paivio, A. (2013). Dual coding theory, word abstractness, and emotion: A critical review of Kousta et al. (2011). *Journal of Experimental Psychology: General, 142*(1), 282–287. https://doi.org/10.1037/a0027004

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development, 76*(4), 763–782. https://doi.org/10.1111/1467-8624.00498-i1

Perfetti, C., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of Functional Literacy* (pp. 189–213). John Benjamins.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.1007/978-3-662-46024-5_8

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science, 21*(2), 1–12. https://doi.org/10.1111/desc.12549

R Developement Core Team. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Article 3.4.3*. https://www.r-project.org/

Rayson, P., Wilson, A., & Leech, G. (2001). Grammatical word class variation within the British National Corpus Sampler. *Language and Computers, 36*(1), 295–306. https://doi.org/10.1163/9789004334113_020

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

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

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language, 35*(1), 185–205. https://doi.org/10.1017/S0305000907008343

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

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences of the United States of America, 112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods, 50*(4), 1568–1580. https://doi.org/10.3758/s13428-017-0981-8

Sandhofer, C., & Smith, L. B. (2007). Learning Adjectives in the Real World: How Learning Nouns Impedes Learning Adjectives. *Language Learning and Development, 3*(3), 233–267. https://doi.org/10.1080/15475440701360465

Savický, P., & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics, 9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The Psychology of Word Meanings* (pp. 223–248). Lawrence Erlbaum Associates.

Snow, C. (2010). Academic language and the challenge of reading for learning about science. *Science, 328*, 450–452.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review, 27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41–78. https://doi.org/10.1207/s15516709cog2901_3

Strömqvist, 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), 45–68. https://doi.org/10.1075/wll.5.1.03str

Ure, J. (1971). Lexical density and register differentiation. In G. E. Perren & J. L. M. Trim (Eds.), Applications of linguistics. *Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969* (pp. 443–452). Cambridge University Press.

van Heuven, W. J. B. Van, Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex, 24*(7), 1767–1777. https://doi.org/10.1093/cercor/bht025

Vigliocco, G., Ponari, M., & Norbury, C. (2018). Learning and Processing Abstract Words and Concepts: Insights From Typical and Atypical Development. *Topics in Cognitive Science, 10*(3), 533–549. https://doi.org/10.1111/tops.12347

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*, 1191–1207. http://macsphere.mcmaster.ca/handle/11375/16227

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science, 24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Weizman, Z. O., & Snow, C. E. (2001). Lexical input 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

Data, Code and Materials Availability Statement

The corpora used in this project are available on the Open Science Framework along with our analysis scripts and Supplementary Materials at https://osf.io/zta29/. Note that for copyright reasons, word order within each text in the picture book corpus has been randomised.

Authorship and Contributorship Statement

ND was involved in conceptualization of the research, led on project management, data analysis, and data curation, and wrote the first draft of the manuscript. YH was involved in conceptualization of the research, provided advice on data analysis, and contributed to reviewing and editing the draft manuscript. AT took the lead on corpus processing and data analysis for some of the measures, and contributed to reviewing and editing the draft manuscript. NB contributed to the acquisition of data, corpus processing, and reviewing and editing the draft manuscript. KN contributed to the conceptualization of the research, and was responsible for funding acquisition, provision of resources, supervision, and reviewing and editing the draft manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This research was funded by The Nuffield Foundation (reference number: EDO 43392). The authors have no conflicts of interest to declare. We would like to thank Nilo Pedrazzini and William Thurwell for their contributions to data processing, and Eleanor Holton, Emma Jackson, Georgia Sandars, Rebecca Williams and Tsvetana Myagkova for their assistance in creating and transcribing the picture book corpus.

Appendices

Appendix A: List of Titles in the Picture Book Corpus

Title	Author	Target age range (years)
A Dog with Nice Ears	Lauren Child	3 to 7
A Great Big Cuddle	Michael Rosen	2 to 7
A Little Bit Brave	Nicola Kinnear	2 to 6
A Squash and a Squeeze	Julia Donaldson	3+
Aliens Love Underpants	Claire Freedman	3+
All the Colours I See	Allegra Agliardi	5+
Along Came A Different	Tom McLaughlin	3+
Animal Stories for 5 year olds	Helen Paiba	5 to 9
Barking for Bagels	Michael Rosen	6+
Bedtime Stories for 5 year olds	Helen Paiba	5 to 9
Brown Bear, Brown Bear, What Do You See?	Bill Jnr Martin	2+
But Excuse Me That is My Book	Lauren Child	4+
Colin and Lee: Carrot and Pea	Morag Hood	3+
Cyril and Pat	Emily Gravett	3 to 7
Dave the Lonely Monster	Anna Kemp	2+
Dear Zoo	Rod Campbell	2+
Dinosaur Roar!	Paul Stickland & Henrietta Stickland	1 to 5
Dogger	Shirley Hughes	2+
Dogs Don't Do Ballet	Anna Kemp & Sara Ogilvie	3+
Duck, Death, and the Tulip	Wolf Erlbruch	4 to 8
Each Peach Pear Plum	Allan Ahlberg & Janet Ahlberg	0+
Elmer	David McKee	3+
FArTHER	Grahame Baker-Smith	7+
Fat Frog	Ruth Miskin	5 to 7
Five Minutes Peace	Jill Murphy	3 to 5
Fox & Goldfish	Nils Pieters	3+
Fox's Socks	Julia Donaldson	1+
Franklin's Flying Bookshop	Jen Campbell	6 to 8
Funny Stories for 5 Year Olds	Helen Paiba	5 to 9
George's Marvellous Medicine	Roald Dahl	7+
Get up!	Ruth Miskin	5 to 7
Giraffe in the Bath and Other Tales	Russell Punter & Lesley Sims	3+
Gracie la Roo Goes to School	Marsha Qualey	6+
Gracie la Roo Sets Sail	Marsha Qualey	5+
Grandad's Island	Benji Davies	5+
Granpa	John Burningham	5 to 7
Guess How Much I Love You	Sam McBratney	2+

Hairy Maclary from Donaldson's Dairy	Lynley Dodd	2+
Hampstead the Hamster	Michael Rosen	5+
Heidi	Johanna Spyri	6+
Hide and Seek	NA	3 to 7
Hide-and-Seek Pig	Julia Donaldson & Axel Scheffler	1+
Hippo has a Hat	Julia Donaldson	0 to 3
Horrid Henry and the Secret Club	Francesca Simon	6 to 11
Horrid Henry tricks the Tooth Fairy	Francesca Simon	7 to 10
Horrid Henry: Ghosts and Ghouls	Francesca Simon	7 to 9
Horrid Henry's Halloween Horrors	Francesca Simon	6 to 11
How to be a Lion	Ed Vere	3+
Hubert Horatio How to Raise your Grown-ups	Lauren Child	7 to 11
I Can Hop	Ruth Miskin	5 to 7
I Want My Hat Back	Jon Klassen	6+
If All the World Were...	Joseph Coelho & Allison Colpoys	0 to 6
In the Bath	Ruth Miskin	5 to 7
Into the Forest	Anthony Browne	8+
Is it a Mermaid?	Candy Gourlay	3 to 7
John Brown, Rose and the Midnight Cat	Jenny Wagner	2 to 4
Joy	Corrinne Averiss	3 to 6
Kitchen Disco	Clare Foges & Al Murphy	5+
Little Beauty	Anthony Browne	2+
Looking for Atlantis	Colin Thompson	8+
Lost and Found	Oliver Jeffers	3+
Loved To Bits	Teresa Heapy & Katie Cleminson	3 to 6
Magical Stories for 5 year olds	Helen Paiba	5 to 9
Me and my Fear	Francesca Sanna	3 to 7
Michael Rosen's Sad Book	Michael Rosen	6+
Mog the Forgetful Cat	Judith Kerr	2+
Monkey Puzzle	Julia Donaldson	3 to 8
Mr Men: Chinese New Year	Adam Hargreaves	3+
Murray the Race Horse	Gavin Puckett	7 to 9
My Father's Arms are a Boat	Stein Erik Lunde	4+
Nice Work for the Cat and the King	Nick Sharratt	6 to 9
Night-Time Cat	Julia Tedd	7
Nip and Chip	Ruth Miskin	5 to 7
No-Bot	Sue Hendra & Paul Linnet	3+
Nog in the Fog	Ruth Miskin	5 to 7
Odd Dog Out	Rob Biddulph	3+
of Thee I sing	Barack Obama	4+
Oi Cat!	Kes Gray	1 to 5

Oi Dog!	Kes & Claire Gray	3+
Oi Frog!	Kes Gray	3+
Oi Goat!	Kes Gray	3+
Owl Babies	Martin Waddell & Patrick Benson	3+
Pants	Giles Andreae	2 to 3
Peace at Last	Jill Murphy	3+
Peck Peck Peck	Lucy Cousins	3 to 5
Peppa goes to London	Lauren Holowaty	3+
Peppa meets Father Christmas	Lauren Holowaty	2 to 6
Peppa the Mermaid	Lauren Holowaty	2 to 6
Peppa's Magical Unicorn	Lauren Holowaty	3+
Princess Mirror-Belle and the Flying Horse	Julia Donaldson	7 to 11
Princess Mirror-Belle and the Sea Monster's Cave	Julia Donaldson	7 to 11
Rabbit & Bear Attack of the Snack	Julian Gough	5 to 7
Rabbit & Bear The Pest in the Nest	Julian Gough	5 to 7
Rabbityness	Jo Empson	5+
Raccoon on the Moon	Russell Punter	3+
Rag the Rat	Ruth Miskin	5 to 7
Red Ned	Ruth Miskin	5 to 7
Room on the Broom	Julia Donaldson	6+
Rosie's Walk	Pat Hutchins	0+
Ruby Red Shoes Goes to London	Kate Knapp	4+
Ruby's Worry	Tom Percival	5+
Run, Run, Run!	Ruth Miskin	5 to 7
Sharing a Shell	Julia Donaldson	2+
Sophie Johnson Unicorn Expert	Morag Hood	3+
Squishy McFluff the Invisible Cat: Seaside Rescue!	Pip Jones	5+
Stardust	Jeanne Willis	5+
Stick Man	Julia Donaldson	6+
Sun Hat Fun	Ruth Miskin	5 to 7
Superworm	Julia Donaldson	2 to 7
Sweep	Louise Greig & Julia Sarda	3+
That's Not my Puppy...	Fiona Watt	0+
That's Not my Unicorn…	Fiona Watt	0+
The Bad-Tempered Ladybird	Eric Carle	2+
The BFG	Roald Dahl	6+
The Building Boy	Ross Montgomery	4+
The Bumblebear	Nadia Shireen	4+
The Cat in the Hat	Dr Seuss	5+
The Day the Crayons Quit	Drew Daywalt & Oliver Jeffers	3 to 7
The Day War Came	Nicola Davies	5+

The Detective Dog	Julia Donaldson	3 to 7
The Flat Rabbit	Bardur Oskarsson	4 to 6
The Gift	Carol Ann Duffy	7+
The Gruffalo	Julia Donaldson	3 to 7
The Gruffalo's Child	Julia Donaldson	3+
The Heart and the Bottle	Oliver Jeffers	6+
The Highway Rat	Julia Donaldson	2 to 6
The Jolly Christmas Postman	Janet Ahlberg & Allan Ahlberg	3 to 5
The Jolly Postman or Other People's Letters	Janet Ahlberg & Allan Ahlberg	3 to 5
The Last Chip: The Story of a Very Hungry Pigeon	Duncan Beedie	3+
The Lion Inside	Rachel Bright	3+
The Marvellous Moon Map	Teresa Heapy & David Litchfield	3 to 7
The Memory Tree	Britta Teckentrup	3 to 5
The Owl who was Afraid of the Dark	Jill Tomlinson	5+
The Paper Dolls	Julia Donaldson	3+
The Pond	Nicola Davies	5 to 7
The Scar	Charlotte Moundlic	5+
The Smartest Giant in Town	Julia Donaldson	4 to 7
The Snail and the Whale	Julia Donaldson	2 to 4
The Storm Whale	Benji Davies	3+
The Storm Whale in Winter	Benji Davies	1+
The Tiger Who Came to Tea	Judith Kerr	2+
The Twits	Roald Dahl	7 to 9
The Ugly Five	Julia Donaldson	2 to 6
The Very Hungry Caterpillar	Eric Carle	0+
The Wonky Donkey	Craig Smith	2 to 6
Tiddler	Julia Donaldson	5 to 11
Tug, tug	Ruth Miskin	5 to 7
Very little Cinderella	Teresa Heapy & Sue Heap	4 to 6
We're Going on a Bear Hunt	Michael Rosen	6+
What Happens Next	Shinsuke Yoshitake	8+
What is Poo?	Katie Daynes	0 to 5
Whatever Next!	Jill Murphy	3 to 5
When Sadness Comes to Call	Eva Eland	3 to 8
Where the Wild Things Are	Maurice Sendak	2+
Where's Spot?	Eric Hill	0+
Willy and the Cloud	Anthony Browne	3 to 7
Willy the Wimp	Anthony Browne	7+
Witchfairy	Brigitte Minne	4+
Zog	Julia Donaldson	2 to 7
Zog and the Flying Doctors	Julia Donaldson	2 to 6

Note. The full corpus is available as .csv files containing word tokens (randomised within each document) on the Open Science Framework project page (https://osf.io/zta29/)

Appendix B: Summary of CHILDES Corpora in the Spoken Language Corpus

Corpus	Child age range	n	Reference
Belfast	2;0-4;5	8	Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. New York: Oxford University Press.
Gathercole/Burns	3;0-6;4	12	Gathercole, V. (1986). The acquisition of the present perfect: explaining differences in the speech of Scottish and American children. *Journal of Child Language*, *13*, 537–560
Howe	1;6-1;8 (session 1) 1;11-2;1 (session 2)	16	Howe, C. (1981). *Acquiring language in a conversational context.* New York: Academic Press.
Korman	6-16 weeks	6	Korman, M., & Lewis, C. (2001). Mothers' and fathers' speech to their infants: Explorations of the complexities of context. In M. Almgren, A. Barreña, M.-J. Ezeizabarrena, I. Idiazaabal, & B. MacWhinney (Eds.), *Research on child language acquisition* (pp. 431-453). Somerville, MA: Cascadilla Press
Lara	1;9-3;3	1	Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1-21. doi:10.1016/j.cogpsych.2017.07.002.
Manchester	1;8-3;0	12	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language, 28*, 127-152.
MPI-EVA Manchester	1;8-3;2	4	Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20* (3), 481-508.
Nuffield	0;11	76	McGillion, M., Pine, J. M., Herbert, J. S., & Matthews, D. (2017). A randomised controlled trial to test the effect of promot-

			ing caregiver contingent talk on language development in infants from diverse socioeconomic status backgrounds. *Journal of Child Psychology and Psychiatry, 58* (10), 1122-1131
Tommerdahl	2;6-3;6	23	Tommerdahl, J. and Kilpatrick, C. (2014). The Reliability of Morphological Analyses in Language Samples. *Journal of Language Testing, 31* (1), 3-18.
Wells	1;6-5;0	32	Wells, C. G. (1981). *Learning through interaction: The study of language development.* Cambridge, UK: Cambridge University Press.

Note. n = number of children in sample. The full corpus is available as .csv files on the Open Science Framework project page (https://osf.io/zta29/)

Appendix C: List of the 50 Book+ Words with Highest Keyness Score and 50 Book– Words with Lowest Keyness Score

Word	Picture book corpus frequency per million	Spoken language corpus frequency per million	Keyness score	Word set
stare	195.31	2.74	16.12	book
voice	293.6	9.96	15.21	book
begin	401.56	18.76	14.31	book
horrid	274.42	10.03	14.2	book
suddenly	252.01	9.74	13.27	book
father	202.7	6.08	13.22	book
everyone	351.18	18.07	12.87	book
yell	130.8	1.48	12.26	book
world	275.73	13.64	12.09	book
giant	290.58	15.5	11.79	book
deep	201.58	8.09	11.7	book
gasp	121.37	1.3	11.62	book
whisper	188.5	7.48	11.36	book
dad	328.98	21.91	10.62	book
leap	129.11	3.17	10.57	book
sigh	114.5	1.91	10.46	book
perfect	168.61	7.45	10.24	book
enormous	116.03	2.47	10.11	book
reply	106.59	2.3	9.48	book
thought	361.99	29.84	9.34	book
shriek	86.13	0.33	9.3	book
mutter	87.35	0.48	9.29	book
large	141.09	6.94	8.92	book
cheer	98.01	2.46	8.67	book
shout	526.77	52.41	8.6	book
dream	175.13	11.63	8.56	book
each	383.27	36.42	8.47	book
towards	141.98	8.12	8.39	book
cave	99.49	3.12	8.35	book
silence	74.84	0.27	8.26	book
sight	106.82	4.52	8.05	book
howl	75.96	0.69	8.04	book
mother	271.95	25.67	7.9	book
ground	231.7	20.58	7.9	book
against	118.92	6.34	7.89	book
breath	105.48	4.67	7.87	book
parent	82.73	1.85	7.83	book
human	73.42	0.7	7.8	book
slowly	157.61	11.54	7.78	book

evening	108.98	5.43	7.71	book
smile	348.68	36.89	7.65	book
hate	99.93	4.38	7.65	book
most	278.81	28.2	7.56	book
street	128.92	8.39	7.56	book
himself	300.22	31.11	7.55	book
peer	66.64	0.31	7.44	book
scream	167.75	14.21	7.34	book
add	114.5	7.37	7.17	book
notice	186.62	17.45	7.16	book
gaze	62.4	0.27	7.05	book
toy	73.3	257.58	0.31	spoken
where	735.23	2459.97	0.3	spoken
train	63.85	237.14	0.3	spoken
because	496.21	1716.97	0.29	spoken
tidy	22.96	102.79	0.29	spoken
cuddle	7.45	49.88	0.29	spoken
hey	67.96	263.26	0.29	spoken
bye	15.41	79.69	0.28	spoken
ooh	24.83	115.13	0.28	spoken
here	701.04	2567.41	0.28	spoken
toilet	8.96	58.76	0.28	spoken
well	820.1	3061.13	0.27	spoken
put	762.26	2919.67	0.26	spoken
tissue	6.86	54.18	0.26	spoken
brick	18.21	99.05	0.26	spoken
yours	36.3	169.13	0.26	spoken
trouser	17.96	99.15	0.26	spoken
do	5333.39	20871.79	0.26	spoken
right	820.27	3235.08	0.26	spoken
giraffe	10.22	69.44	0.25	spoken
oops	5.37	51.39	0.25	spoken
doll	32.06	160.44	0.25	spoken
we	1391.55	5698.67	0.25	spoken
today	110.42	482.4	0.24	spoken
what	2794.66	11487.89	0.24	spoken
yum	12.12	81.21	0.24	spoken
naughty	37.44	187.53	0.24	spoken
yesterday	27.44	150.82	0.23	spoken
car	90.86	425.88	0.23	spoken
oy	3.25	50.83	0.22	spoken
whee	6.69	67.96	0.21	spoken
you	7427	34770.54	0.21	spoken

want	712.68	3370.12	0.21	spoken
penguin	4.18	57.24	0.21	spoken
yes	515.37	2786.8	0.19	spoken
nursery	3.3	64.39	0.18	spoken
poorly	3.45	66.25	0.18	spoken
shall	204.17	1366.71	0.16	spoken
careful	35.75	303.12	0.15	spoken
mm	6.21	106.22	0.14	spoken
jigsaw	5.45	104.19	0.14	spoken
wee	16.98	246.49	0.11	spoken
hm	46.59	547.67	0.1	spoken
oh	828.94	8781.37	0.1	spoken
whoops	4.35	170.3	0.08	spoken
okay	89.43	1581.96	0.06	spoken
pardon	18.23	469.11	0.06	spoken
darling	20.93	629.48	0.05	spoken
alright	4.49	519.47	0.03	spoken
yeah	28.24	2554.14	0.01	spoken

Note. Frequency columns show average reduced frequencies per million prior to the addition of the constant (10)

License