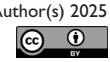
Article

FIRST LANGUAGE

Charting the frequency and diversity of emotion words in children's language: Written language matters

First Language
2025, Vol. 45(4) 457–475



DOI: 10.1177/01427237251339788
journals.sagepub.com/home/fla
Sage

Yuzhen Dong and Kate Nation
Department of Experimental Psychology, University of Oxford, UK

Abstract

Emotion words allow us to identify, describe and regulate our emotional states. Emotion vocabulary grows through childhood, but little research has considered emotion words in the context of children's written language. To address this gap, we used a cross-corpus developmental approach to chart the emergence of emotion words in children's reading experience and in their own writing. For comparison, we also captured occurrences of the same set of emotion words in age-matched samples of children's spoken language experience via caregiver child-directed speech and television programmes. We observed that even books targeted at preschoolers for shared reading contained more unique emotion words than both caregiver speech and television language. As the targeted age of books increased through mid-childhood and early adolescence, the frequency and diversity of emotion words increased further. This pattern was also seen in children's own writing, with more unique and diverse emotion words being used by older children. These findings indicate that written language requires children to comprehend and produce emotion words that are rare in everyday conversations. We speculate that this linguistic experience may play a role in emotional development by providing opportunities to consider and communicate mental situations beyond the everyday.

Keywords

Language, emotion, development, reading, childhood

Corresponding author:
Yuzhen Dong, Department of Experimental Psychology, Anna Watts Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK.
Email: yuzhen.dong@psy.ox.ac.uk

Introduction

Language is considered a 'key ingredient' in the development of social–emotional understanding and empathy (Lindquist, 2017). Language offers the tools for identifying and describing our own emotions and those of others. From the early stages of development, children's linguistic experiences are deeply embedded in social interaction, setting the stage for their burgeoning emotional understanding later on (Dunn et al., 1991). There is good evidence to show that emotion words, or affective labels, play a fundamental role in emotional development. For example, the breadth of preschoolers' emotion vocabulary has been linked to their emotion regulation strategies (Streubel et al., 2020), and more generally, the capacity to label and differentiate emotions has implications for mental well-being through childhood, adolescence and adulthood (e.g. Nook, 2021; Torre & Lieberman, 2018). Most studies on the development of emotion vocabulary have focused on infancy and early childhood, and children's experience with emotional language has been captured via child-directed speech, typically sampling the language of caregivers as they interact with their children (e.g. Nencheva et al., 2023; Ogren & Sandhofer, 2021). There has been less emphasis on mid-childhood, yet Baron-Cohen et al. (2010) found that the size of children's emotion word vocabulary doubles every 2 years between 4 and 11 years of age. Interestingly, this expansion in emotion word vocabulary aligns with the onset of literacy. Children's books contain more words and more diverse and complex words than child-directed speech (e.g. Dawson et al., 2021) and as children learn new vocabulary via reading (e.g. Joseph et al., 2014), it is likely that emotion words, especially more complex emotion words, might be acquired via experience with written language. This feels like a plausible hypothesis, but the evidence base lacks information about the nature of emotion vocabulary afforded by written language input. This is a critical gap in our understanding of how emotional development might be influenced by written language. We address this gap in two ways. First, we chart the occurrence of emotion words in books targeted at children of different ages and second, we consider children's usage of emotion words in their own writing in cross-sectional analyses of stories written by 5- to 13-year olds. Our aim is therefore to describe the occurrence and usage of emotion words in language written *for* children and in the language written *by* children themselves. We also ask how this compares with the occurrence of the same set of emotion words in child-directed speech, and in the language of television shows targeted at children of the same age.

According to the constructionist hypothesis of emotional development, language supports the acquisition of emotional knowledge and is considered critical to the development of different aspects of emotional understanding (Barrett, 2017; Hoemann et al., 2019; Lindquist, 2017; Shablack & Lindquist, 2019). Language is situated in social interaction from infancy onwards, and this is reflected in the association between children's experience with emotional language and their subsequent emotional understanding. For example, Dunn et al. (1991) coded 'feeling state' language (e.g. happy) from recordings of everyday mother–child interactions at 36 months of age. They noted that children who experienced more feeling state language at 36 months were better able to understand others' emotions at 6 years of age. In line with the view that language is critical to the construction of emotion, there is also an association between language competence and

emotion understanding in preschoolers (e.g. Widen & Russell, 2003, 2008) and older children (e.g. Beck et al., 2012; Griffiths et al., 2020).

Research examining how young children experience and use emotion words within their language environment has provided valuable insights into the linguistic underpinnings of emotional development. Contributing to this, Ogren and Sandhofer (2021) analysed language transcripts from the Child Language Data Exchange System (CHILDES) database for children aged 15 to 47 months. They described the characteristics of young children's emotion lexicon, including the frequency and variety of emotion words and factors influencing individual differences in their production. Nencheva et al. (2023) used data from Wordbank (Frank et al., 2017; infant receptive and expressive vocabulary as estimated by the MacArthur-Bates Communicative Development Inventory parental checklist) and CHILDES (caregiver speech) to study the relationships between toddlers' language production and caregiver input. They found that toddlers develop a broad emotional vocabulary early in life, influenced by a network of semantically connected words. They also showed in a longitudinal analysis that how caregivers use emotion and mental state labels predicts children's emotion label production over time.

While these studies document important links between language input and the development of emotional vocabulary in children, the field has primarily focused on the spoken language environment, especially within the context of child-directed speech and early childhood interactions. Yet, emotional vocabulary continues to develop beyond early childhood and might therefore be facilitated by the onset of literacy. Baron-Cohen et al. (2010) used parent/teacher checklist data to estimate whether 4- to 16-year olds understand 336 emotion words. They found a general increase in emotion word vocabulary by age, and they noted that emotion word vocabulary doubles every 2 years between 4 and 11 years of age. In line with this observation, Ponari et al. (2018) analysed the age of acquisition ratings for over 13,000 words and identified 8.5 years as the age, which shows the steepest increase in acquisition of abstract words, including words describing emotions. These findings are certainly consistent with the idea that written language provides critical language input that might help drive the acquisition of emotion vocabulary. Reviewing the nature of children's 'book language', Nation et al. (2022) argued that exposure to written language provides children with opportunities to experience language that differs from everyday usage, including the complex and nuanced language associated with emotions and mental states. Even books written for preschoolers contain more linguistically complex words than child-directed speech (Dawson et al., 2021; Montag et al., 2015). For older children, Korochkina et al. (2024) found that 28% of words in books for 7- to 9-year olds never appear in age-appropriate television programmes. In addition, Dawson et al. (2021) noted that books for preschoolers contain more emotionally arousing words than child-directed speech. Readers need to navigate book language to understand social relationships and make inferences about the mental state of characters (Kim et al., 2015; Mar et al., 2006; Siller et al., 2014; White et al., 2009), and children's story books therefore provide opportunities to adopt the emotional perspectives of different characters and to understand empathy (e.g. Hogan, 2011; Kucirkova, 2019). Being read to, joint reading and independent reading are all associated with children's social, emotional and educational outcomes (Green et al., 2023). From this view, the development of emotion word vocabulary in mid-childhood might be a

result of experience with books and reading. Thus, as children's language experiences evolve with the onset of literacy, there is a clear need to investigate how written language provides opportunities for children to experience emotion words.

One way to measure children's understanding and use of emotion words is to elicit them in production tasks. Several studies have reported that children's production increases through childhood and becomes more nuanced and multidimensional over time (e.g. Grosse et al., 2021; Nook et al., 2020). For example, emotion word labelling of facial expressions increases with age between 3 and 5 years (Widen & Russell, 2008). Using a vignette test which depicts a child protagonist in a typical emotion-eliciting situation, Grosse et al. (2021) asked 123 four- to eleven-year-old children to name the emotion the protagonist in the vignette might feel. They found that older children produced more emotion words, and their usage became more adult-like. Using a word generation task (e.g. 'think about a child who is feeling sad and then write down as many words as you can to describe this feeling'), Doost et al. (1999) found that older children in secondary schools produced significantly more emotion words than the younger children in primary schools. While these findings mirror the increases in emotion word comprehension through mid-childhood reported by Baron-Cohen et al. (2010), it is notable that existing studies have all probed emotion word usage directly, using simple word generation or picture prompts. Lacking is a more nuanced and naturalistic perspective on children's use of emotion words.

In summary, the importance of emotion words in emotional development is well accepted. However, what is lacking are data that consider emotion vocabulary in children's written language experience, both in terms of what children read and what they produce themselves. We took a corpus linguistics approach to address these two evidence gaps, using data from several corpora that sampled language targeted at children of different ages via child-directed speech, television shows and books, and one corpus of stories written by children. Across each corpus, we considered the 336 emotion words identified by Baron-Cohen et al. (2010) and described the frequency and diversity of these emotion words. We predicted that children's books would contain a more extensive and diverse range of emotion words than child-directed speech and television, and we also expected language written for older children to contain more diverse use of emotion words than younger children, and for this pattern to also be seen in children's own writing.

Methods

We used six children's language corpora (Table 1) and investigated the relative occurrence of 336 emotion words, as selected by Baron-Cohen et al. (2010). They defined an emotion word as one that describes a mental state with an emotional dimension (i.e. it could be preceded by 'I feel x' or 'he/she looks x'). They also included mental state terms that were epistemic but had an unambiguous emotional dimension. Excluded from their list were words denoting mental states that could be interpreted as purely bodily states, epistemic states without an emotional dimension, slang and swear words, and words deemed too advanced for the age groups surveyed. Baron-Cohen et al. (2010) estimated the development of children's comprehension using a checklist procedure in which parents and teachers indicated whether children and adolescents (aged 4–16 years) would

Table 1. Description of the six corpora used in this paper.

Corpus name	Source	Targeted age range	Size (Type)	Size (Tokens)
Child-directed speech	CHILDES-U.K. (MacWhinney, 2000)	0–6 years	24,129	3,853,976
Picture books	ReadOxford (Dawson et al., 2021)	0–7 years	11,561	319,435
CBeebies (TV)	SUBTLEX-U.K. (van Heuven et al., 2014)	0–6 years	27,236	5,848,083
CBBC (TV)	SUBTLEX-U.K. (van Heuven et al., 2014)	6–12 years	58,691	13,612,278
Reading books	Oxford Children's Language Corpus – 2019 Reading (OUP[1])	5–16 years	377,108	63,494,697
Children's writing	Oxford Children's Language Corpus – 2019 Writing (OUP)	5–13 years	223,594	47,841,388

Note. OUP = Oxford University Press.

understand each of the 336 words. Most of the words ($N = 317$) are included in the valence and arousal ratings provided by Warriner et al. (2013). In language, valence refers to the pleasantness of a word and the extent of its positivity or negativity, and arousal refers to the intensity of emotion provoked by a word. These ratings are measured on a scale from 1 to 9 (where 1 represents *very unhappy/not arousing*, and 9 represents very happy/arousing). The mean valence score of the 317 words is 4.83 ($SD = 1.92$) and the mean arousal score is 4.65 ($SD = 1.00$). Those words ($N = 19$) not found in Warriner et al.'s list included affixed forms such as *discontented* and *mistreated*, compound words like *fed-up* and *bad-tempered*, and low-frequency words such as *commiserating*.

The 336 words are reproduced in Supplemental Appendix 1, along with item-level summary statistics from our analyses.

Children's language corpora

Child-directed speech was generated from 10 corpora in the English-U.K. section of the CHILDES database (MacWhinney, 2000), following Dawson et al. (2021). The sample comprised all suitable corpora from this collection, except for those that focused on specific populations (e.g. children with language impairments). The final set of 10 corpora (see Appendix B in Dawson et al., 2021, for a summary) 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. Nine of the 10 corpora comprised recordings of unstructured interactions. Of these, eight were collected in home environments during naturalistic routines such as mealtimes and bedtimes as well as free play sessions and one corpus was recorded in a university laboratory playroom, also involving unstructured play. The other corpus

was recorded in a preschool nursery setting and captured more structured task-based interactions such as block construction and art. 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 (24,129 unique). Most of the child-directed speech (77.83% of tokens) was produced by mothers, while 10.24% was produced by investigators. See Supplemental Appendix 2 for the age distribution of this sample of child-directed speech.

The *Picture Book corpus* comprised 160 children's fiction books, selected to be representative of the type of reading material children encounter in shared reading contexts in the United Kingdom (Dawson et al., 2021). Of the 160 picture books, 98 were labelled as narrative (61.25%), 61 as rhyming (38.13%) and one was unlabelled. The corpus captures book language input that is age-matched to the child-directed speech documented in CHILDES. It was created by Dawson et al. (2021) by first generating a list of book titles with a target age range of 0 to 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 a combination of retailer bestseller lists and recommendations from literacy charities, book review sites and teachers. The dataset contains 319,435 words and 11,561 types, and the distribution of the target age of the books is in Supplemental Appendix 2.

The *Reading Book corpus* was sampled from the reading component of the Oxford Children's Language Corpus developed and held by Oxford University Press (OUP, Wild et al., 2013). It was initiated in 2006 to guide the preparation of children's dictionaries. The corpus contains a wide range of material that children encounter during their reading experience, including classic and modern fiction, non-fiction, textbooks, websites and magazines. We accessed the 2019 version of the corpus, totalling 63,494,697 words (377,108 types) written for 5- to 16-year-old children. Frequency lists of all the words in the corpus were downloaded from Sketch Engine (http://www.sketchengine.eu), with the kind permission of Oxford University Press. Some documents (45.8%) were tagged with the targeted Key Stage. This refers to bandings within the education system of England and Wales, with 5- to 7-year olds falling into Key Stage 1, 7- to 11-year olds into Key Stage 2, 11- to 14-year olds into Key Stage 3 and 14- to 16-year olds into Key Stage 4. Books targeting children at Key Stage 2 were most prevalent (18 million tokens), followed by Key Stage 3 (14 million tokens). Books targeting younger children (Key Stage 1) and older children (Key Stage 4) are smaller, each comprising 2 million tokens.

CBeebies and *CBBC* data were extracted from the children's sub-corpora of SUBTLEX-U.K. (van Heuven et al., 2014), which is a corpus of subtitles from television programmes shown by the British Broadcasting Corporation (BBC). CBBC content is targeted at children aged 6 to 12 years, while its sister channel, CBeebies, is aimed at preschool children. The CBeebies norms were derived from 5,848,083 words and contain 27,236 types, while the CBBC norms were derived from 13,612,278 words and contain 58,691 types, as reported in van Heuven et al. (2014). We used these data as a proxy for spoken language input, but note that the discourse and narrative context of television language is quite different from the type of spoken language experienced in day-to-day conversation, as indexed by CHILDES and child-directed speech. We return to consider this further in the Discussion.

The *Writing Corpus* was sampled from the writing component of the Oxford Children's Language Corpus developed and held by Oxford University Press (Wild et al., 2013). This is a dynamic corpus that contains stories submitted as part of the BBC 500 Words annual writing competition for children. This has been running for over 10 years with the same format in which children (aged 5–13 years until 2020, and then aged 5–11 when the competition returned in 2023) across the United Kingdom were invited to submit a story on any theme or topic, so long as the word count was not greater than 500 words. We selected all stories submitted in 2019 ($N = 107{,}273$; approximately 55 million word tokens; see also Dong et al., 2024; Hsiao et al., 2023, 2024). Each story was tagged with the Key Stage of the child author. Most entries (59%) came from children in Key Stage 2; 39% of entries came from children in Key Stage 3 and only 2% from the youngest children in Key Stage 1. More girls (59.44%) contributed stories than boys (40.56%).

Analysis procedure

We analysed the distribution and diversity of the 336 emotion words identified by Baron-Cohen et al. (2010) across each corpus and our results are organised into three short sections. First, we considered the language directed at preschool children and compare emotion word input in picture books, child-directed speech and television subtitles from CBeebies. We then considered language targeted at older children and compared emotion vocabulary in children's reading books with television subtitles from CBBC. Finally, we considered children's own production by analysing the occurrence of emotion words in their own written stories. Data and code associated with this paper are available on the Open Science Framework website (https://osf.io/hc7yk/).

Given the size of the datasets and the potential issue of overpowering (Egbert et al., 2022), inferential statistics are not necessarily helpful. We thus describe and interpret patterns based on cross- and within-corpus visualisations, supplemented with inferential statistics where appropriate. To visualise the distribution of emotion words across the corpora, we calculated the number of occurrences of each emotion word in each corpus and plotted the frequency distribution.

To compare the frequency distributions of emotion words across corpora, we first computed the Zipfian slopes for each corpus to capture the rank-frequency profiles. For each slope, we performed a bootstrapping procedure with 1000 resamples to estimate the mean and 95% confidence intervals. During each iteration, frequency data were randomly sampled with replacement. The Zipfian slope was recalculated using a discrete power-law model using the *poweRlaw* package in R, and the slope parameter was estimated via maximum likelihood estimation. A larger value for Zipfian slope indicates a flatter slope and more low-frequency words. We then compared the distributions of bootstrapped Zipfian slopes between corpora using the Mann–Whitney *U* test. This is a non-parametric test that does not assume normality and it is appropriate for comparing rank-based differences between distributions. This test was applied to determine whether there was a significant difference in Zipfian slopes between corpora.

To compare the diversity of emotion words between corpora of different sizes, we adopted the sampling methods used by Montag et al. (2015) and Dawson et al. (2021). We first shuffled all the words in a given corpus across texts and documents, and then

extracted multiple random samples from each corpus with replacement. The random samples ranged in size from 100 to 50,000 words, increasing in increments of 100 words each time. One hundred samples were generated at each sample size, each based on a new random sample. For each sample of tokens, we counted the occurrence of the 336 emotion words. We then aggregated and plotted the number of emotion word types at each sample size. To assess whether there were significant differences in the number of emotion word types across corpora, we performed a one-way ANOVA at two sample sizes (25,000 and 50,000 tokens). At each sample size, 100 samples were generated for each corpus for analysis.

Results

Language directed at preschool children

Figure 1 (panels a–c) shows the overall distribution of the emotion words in the three corpora (child-directed speech, CBeebies, picture books). The *x*-axis is the frequency rank of individual words from most frequent (left) to least frequent (right). The *y*-axis is the raw frequency of the emotion word. We see that all distributions are Zipfian, where the frequency of a word is inversely proportional to its frequency rank. We compared the frequency distributions of the corpora using bootstrapped Zipfian slopes and a Mann–Whitney *U* test. The mean Zipfian slope was 1.51 (95% CI [1.37, 1.81]) for child-directed speech, 1.70 (95% CI [1.37, 2.04]) for the CBeebies television subtitles and 2.10 (95% CI [1.44, 4.02]) for the picture book corpus. This shows that the picture book corpus has a larger value for the Zipfian slope, and therefore has a flatter slope and contains more low-frequency words than CBeebies, followed by child-directed speech. Pairwise comparisons using Mann–Whitney *U* revealed significant differences between each pair of corpora (CBeebies vs. child-directed speech: $U = 265{,}625$, $p < .001$, rank-biserial correlation = 0.47; picture books vs. CBeebies: $U = 330875$, $p < .001$, rank-biserial correlation = 0.33; picture books vs. child-directed speech: $U = 174{,}436$, $p < .001$, rank-biserial correlation = 0.65). It is important to note that the power-law model did not provide a perfect fit for the frequency distributions of the picture books corpus, which is the smallest corpus. To complement these statistical analyses, visual inspection of the distributions shows that spoken language in child-directed speech and children's television programmes has a steeper gradient and a longer tail, while written language in children's picture books is more spread.

Figure 2 (panel a) shows the mean number of emotion word types at each sample size in the same three corpora, namely child-directed speech, CBeebies and picture books. The data show that at each sample size, the picture book corpus contains a greater number of unique emotion word types than the CBeebies subtitles, followed by child-directed speech. Differences also emerge in the slopes of the lines. The number of emotion word types in picture books shows a greater increase in unique emotion word types per unit increase in word tokens than both CBeebies and child-directed speech. One-way ANOVA shows that the three corpora differ significantly, both when sample size was $N = 25{,}000$: $F(2, 297) = 1758$, $p < .001$; and at sample size $N = 50{,}000$: $F(2, 297) = 2030$, $p < .001$. Post hoc pairwise comparisons using Bonferroni adjustment found significant differences between each pair of corpora, supporting the observation that the picture book

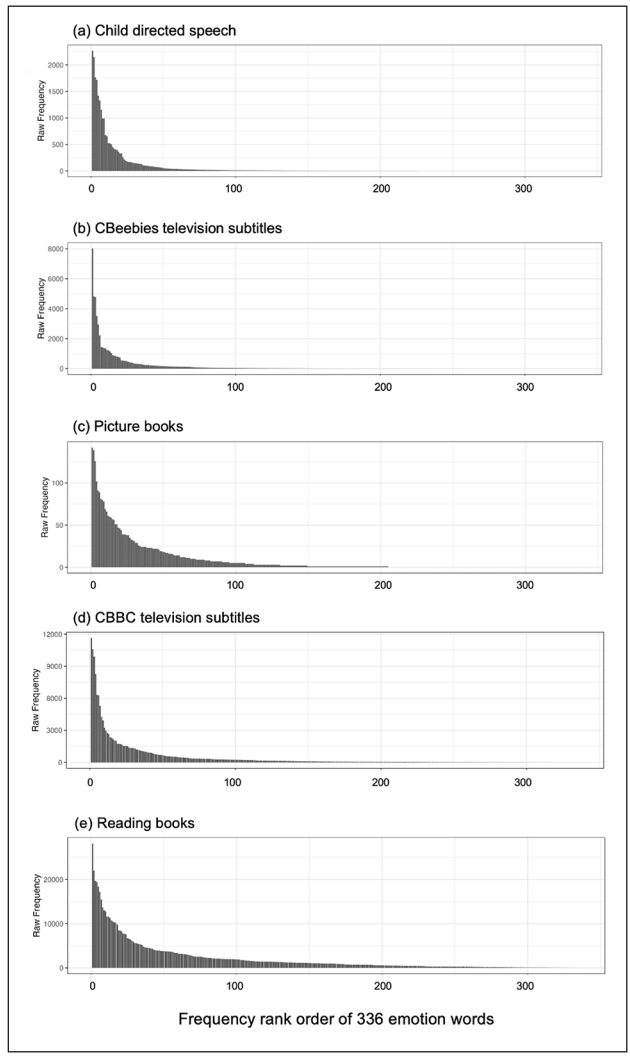


Figure 1. Frequency of emotion words in language for younger children (a–c) and older children (d and e).

corpus contains a greater diversity of emotion word types than the CBeebies subtitles which in turn show more diversity in emotion word types than child-directed speech.

Language directed at older children

Figure 1 (panels d and e) shows the frequency distributions of emotion words in the CBBC and reading book corpora, which once again are clearly Zipfian. We used the same procedure to compare the frequency distributions across the two corpora using

bootstrapped Zipfian slopes and a Mann–Whitney *U* test. The mean Zipfian slope was 1.92 (95% CI [1.57, 2.56]) for the CBBC television subtitles and 2.24 (95% CI [1.87, 2.79]) for the reading book corpus. The Mann–Whitney *U* test revealed significant differences between the two distributions, $U = 172{,}191$, $p < .001$, rank-biserial correlation = 0.66. Emotion words in CBBC television subtitles show a steeper gradient and a longer tail, while written language is more spread, with less concentration of high-frequency words.

Turning to the lexical diversity of the two registers, Figure 2 (panel b) shows that at any given sample size, books contain a greater number of unique emotion word types than CBBC television subtitles. The steeper gradient for book language indicates a greater increase in unique word types per unit increase in word tokens. One-way ANOVA shows that the two corpora differ significantly at each of the two sample sizes tested, at $N = 25{,}000$: $F(2, 198) = 1161$, $p < .001$; and at $N = 50{,}000$: $F(2, 198) = 2879$, $p < .001$.

To model lexical diversity across targeted age, we used the Key Stage metadata available for around half of the reading books. We conducted sampling at each Key Stage to calculate the unique number of emotion word types at random samples of each corpus. Figure 3 (panel a) shows that at any given sample size, books targeted at older children contain a greater number of unique emotion word types than books targeted at younger children. One-way ANOVA shows that the number of emotion word types differ significantly for different Key Stages in the reading corpus at each of the two sample sizes tested, at $N = 25{,}000$: $F(2, 396) = 481.8$, $p < .001$; and at $N = 50{,}000$: $F(2, 396) = 711.3$, $p < .001$. Post hoc pairwise comparisons using Bonferroni adjustment found significant differences between each of the consecutive Key Stages in the reading book corpus (KS2-KS3: $p = .001$, all other ps $< .001$), supporting the observation that books targeted at older children contain more unique emotion word types than books targeted at younger children.

Language written by children

Each story in the children's writing corpus was tagged by the Key Stage of the child author. Most entries (59%) came from children in Key Stage 2; 39% of entries came from children in Key Stage 3 and only 2% from the youngest children in Key Stage 1. We conducted sampling at each Key Stage to calculate the unique number of emotion word types in random samples of each reading corpus. Figure 3 (panel b) shows that at any given sample size, stories written by older children contain a greater diversity of emotion words than stories written by younger children. One-way ANOVA shows that the number of emotion word types differ significantly for different Key Stages in the writing corpus at each of the two sample sizes tested, at $N = 25{,}000$: $F(2, 297) = 469.8$, $p < .001$; $N = 50{,}000$: $F(2, 297) = 817.4$, $p < .001$. Post hoc pairwise comparisons using Bonferroni adjustment found significant differences between each of the consecutive Key Stages in the writing corpus (all ps $< .001$), supporting the observation that stories written by older children contain more unique emotion word types than stories written by younger children.

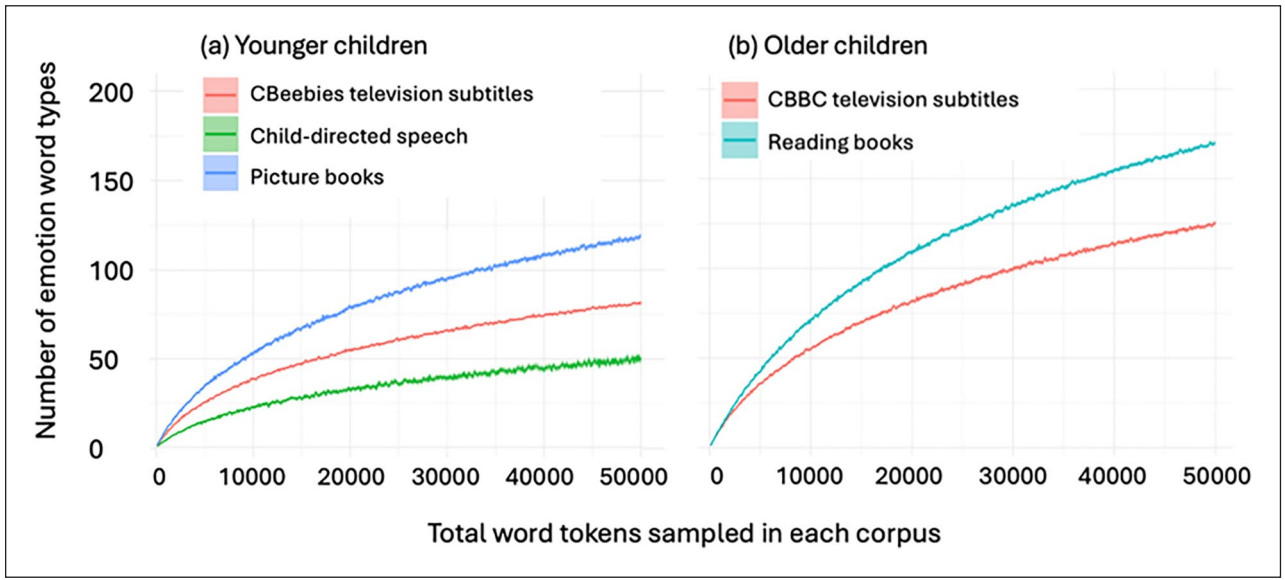


Figure 2. Diversity of emotion words in language for (a) younger children and (b) older children.
Note. The shaded area around the line represents the 95% confidence interval around the mean at each sample size. The intervals are not easily visible because they are extremely narrow, reflecting low variability in the sampled data.

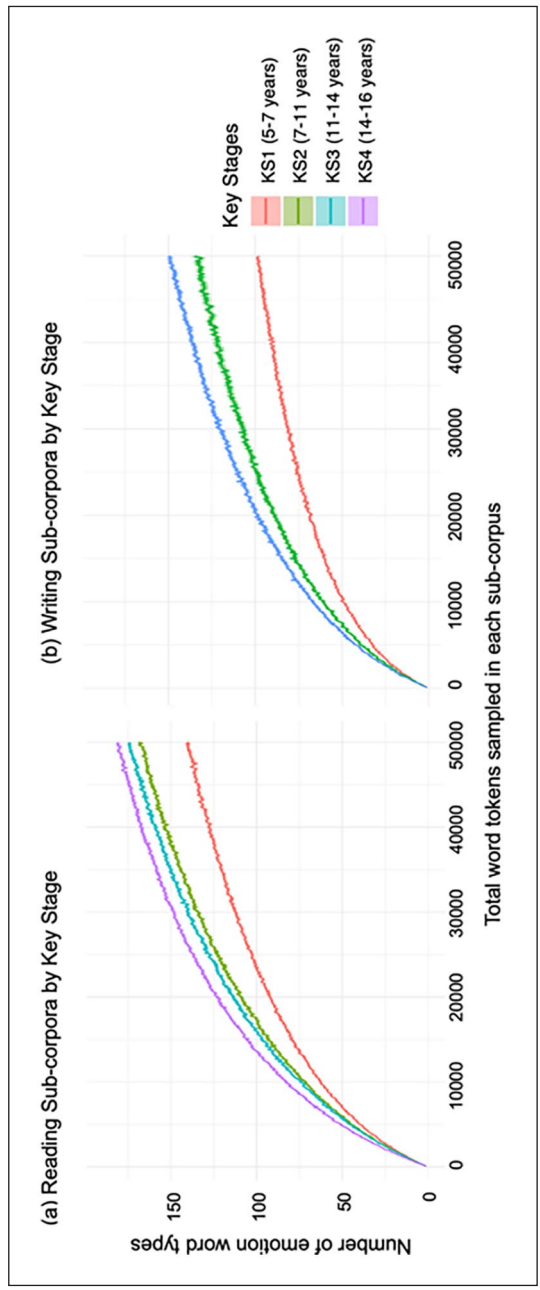


Figure 3. Diversity of emotion words in children's (a) reading and (b) writing by Key Stage.
Note. The shaded area around the line represents the 95% confidence interval around the mean at that sample size. The intervals are not easily visible because they are extremely narrow, reflecting low variability in the sampled data.

Discussion

Our study charted emotion word frequency and diversity across children's language, with a focus on written language. Comparing language input afforded by the language children hear with the language they read (and hear in the context of shared reading), our findings show a greater diversity of emotion words in book language than spoken language. The frequency distribution of emotion words was Zipfian, mirroring the general distribution of word frequencies in language, where a small set of words appear very frequently, with the majority forming the long tail of the distribution (Piantadosi, 2014). As compared to spoken language, the frequency distribution for emotion words in books is more spread, suggesting that children are more likely to encounter low-frequency emotion words in books rather than via child-directed speech or from watching television. In language targeted at preschool children, picture books contained a greater diversity of emotion words than the language of CBeebies television programmes, which in turn, contained more diversity than child-directed speech. Similar patterns were observed for older children, where books for independent reading contained more emotion words than the CBBC television programmes. We were also able to compare books by Key Stage, and this showed that books written for older children contain a wider range of emotion words relative to those written for younger children. Mirroring this, older children used more emotion words in their own writing, relative to the stories written by younger children.

At a general level, our findings are consistent with other investigations of children's book language that document greater linguistic complexity and diversity in books than child-directed speech (Dawson et al., 2021, 2023; Hsiao et al., 2023). These differences can be attributed to the inherent characteristics of written versus spoken language. Written language, being more decontextualised, requires readers to rely on the text itself to understand the intended meaning, without the aid of paralinguistic cues and a shared situation. This necessitates lexical diversity and greater linguistic complexity more generally. Clearly, spoken language interactions provide rich input from which children can learn about emotion via shared context and multimodal cues such as prosody, facial expression and gestures. When reading, however, the greater diversity of emotion words in written language exposes children to a wider range of nuanced emotion words and words that are rarely encountered in everyday conversations.

Extending beyond previous work that has compared lexical diversity in children's books versus child-directed speech, our study included comparisons with television language. We found that both CBeebies and CBBC programmes contained a greater range of emotion words than child-directed speech, but both had less diversity than books. This middling position fits with the observation that while television programmes are often scripted, they also contain elements of spontaneous speech (Zhang & Gu, 2023). Similar findings have been reported for other types of media language, such as the language used in educational apps for preschoolers. Kolak et al. (2023) found these contained lower frequency words than child-directed speech, yet shorter utterances than books. Gowenlock et al. (2024) reviewed language exposure to video content and its impact on linguistic development in children aged 3 to 11 years, and they reported mixed results

depending on video quality and viewing context. Looking across the different input corpora in our study, our findings suggest that engaging with books allows children to encounter more diverse language and language experiences that are distinct from everyday speech and other types of media, including emotional language. Books, therefore, offer children the chance to encounter language that portrays emotion and mental states and in turn, this might support the development of a more comprehensive understanding of emotional concepts.

Written language aimed at older children contained a more diverse range of emotion words than books written for younger children. In particular, emotion word diversity began to increase for 7- to 11-year olds, as estimated by Key Stage 2 books. This was reflected in the children's own writing. Older children produced more emotion words, and the difference was the largest between Key Stage 1 and Key Stage 2. This echoes with Baron-Cohen et al.'s (2010) parent checklist data, which estimated a rapid expansion in emotion word vocabulary at about 8.5 years. We suggest that as children enter Key Stage 2 and reading expertise expands, books offer an opportunity to experience rich and nuanced language, including words that articulate emotions and mental states. This experience might then serve to enrich children's understanding of emotion concepts, such that the increased intake of emotional language is also reflected in how children use emotion words in their own writing. This is supported by the two similar looking graphs in Figure 3. Indeed, the act of crafting writing might itself be a driver for developmental change. On this view, the need to produce written language in such a way that it conveys the intended meaning for the reader might extend children's use of emotion language to communicate empathy and emotional states, and in turn, this will be reflected in greater understanding of empathy and emotional states (see Dong et al., 2024, for further discussion). We are not able to forge direct links across the two corpora, and so this speculation awaits testing experimentally, and longitudinal data would be particularly valuable. It is also important to note that using an emotion word does not guarantee that a child understands or can express the associated emotion, and vice versa. Previous research by Ogren and Sandhofer (2021) and Nencheva et al. (2023) studied how children's spoken production of emotion and mental state labels is scaffolded by caregivers' use of these labels in early childhood. Future research could extend these approaches to older children and to written language, assessing how a child's written and spoken language experiences relate to their comprehension and production of emotion words in both speech and in writing. Findings from such a study could provide valuable and more direct insights into how language experience shapes emotional development, and how language experience provides the linguistic proficiency needed to express nuances in emotion.

Our intention was to document the occurrence of emotion words by corpus and from this make inferences about when children experience (and produce) emotion vocabulary. It is important to note that while this 'bag-of-words' approach captures type and token frequency and reveals patterns that can inform psychological theory (e.g. Jackson et al., 2022), it does not capture the language environment in which words occur. This is important as contextual factors such as negation and figurative usage (e.g. metaphors and idiomatic expressions) can alter the meaning of emotion words. Linguistic context in written language is important, not least because of the decontextualised nature of

reading. In child-directed speech and shared reading, for instance, adults might describe emotions about specific incidents, events or narratives. This interaction does more than introduce children to emotion words; it contextualises them, and potentially enriches children's understanding of emotional concepts. The language used by adults to describe these situations often mirrors the valence of the emotion words themselves, providing a comprehensive emotional landscape (e.g. Nencheva et al., 2023). For example, words accompanying a description of a joyful event are likely to be positively valenced, thereby reinforcing the emotion concept of happiness. Moreover, the diversity of word types in children's language input is not just related to the amount of language that children are exposed to in different modalities, but is also associated with variables such as the size of caregivers' vocabularies (Green et al., 2023; Montag et al., 2015). While children surely build from this foundation as they read, the act of independent reading will require them to access new language from the books directly, and to do so without the scaffolding support of a caregiver or teacher. The richness and complexity of book language provides opportunities to learn this language, and future work should harness the power of large language models operating over different types of language corpora to take a data-driven approach to understanding emotion language and its emergence in language written for and by children, beyond investigation of a set of words preselected as being 'emotion words'.

Our results need to be considered in the context of the characteristics and content of each of the corpora, and the cultural and interpersonal context in which they were sampled. For example, most of the child-directed speech data sampled from CHILDES reflect the language children hear in the home environment. While comprehensive, this does not capture language input from other preschool settings or other children. Even within the home environment, different contexts might afford different levels and complexities of emotion talk across episodes. For example, a recording during a sibling argument is likely to contain different emotion words than a mealtime conversation. Cultural differences might also influence how emotion words are used in speech and different contexts. Similarly, there are different genres of written language with sub-genres nested beneath fiction and non-fiction, and differences in language content and structure are to be expected. For instance, narrative fiction contains a larger proportion of complex emotion words (e.g. amusement, despair) and words used in an emotive sense (e.g. a sigh of relief vs. hurricane relief fund) than non-fiction (Schwering et al., 2021). Our corpora comprised mainly fiction, and there was insufficient metadata to investigate usage at a finer grain size.

Before closing, we should also note that differences between corpora might introduce unintended confounds. For example, while both the child-directed speech corpus and the picture book corpus were chosen as they targeted preschool-aged children of the same age, the age distributions of tokens within the two corpora differ (see Supplemental Appendix 2). The picture book corpus skews toward older children, with the majority of tokens aimed at 4-year olds and above. This is understandable, as picture books written for older children are generally longer and therefore contain more words. As for all corpus analyses of this type, it is impossible to determine precisely how, when and whether children actually experience the language described in a corpus. Nevertheless, our study represents a valuable first step in identifying trends in children's language exposure.

Future research could take recordings from a variety of settings beyond the home (e.g. classrooms), and children's exposure to print materials could be documented more systematically via digital tools or reading diaries. Future studies will be better placed to provide a more nuanced and accurate understanding of emotion word occurrence and usage across contexts and targeted ages. Ultimately, it would be interesting for tracking to happen at an individual level and for this individual-level exposure to be related to the child's learning and language processing.

In conclusion, while previous work has considered emotion word development, the potential role of experience with book language has not been considered directly. This is unfortunate as written language has discourse properties that might provide ideal opportunities for children to experience and therefore learn emotion words, and perhaps nuances in emotion concept too. To start to fill this gap, we charted the frequency and diversity of emotion words across different language corpora documenting children's books, child-directed speech and television media, as well as the emotion words children use in their own written stories. We found a consistent pattern of more diverse emotion words in children's books. This is consistent with children encountering more complex emotion words and mental state labels via reading, which are subsequently reflected in their own writing.

Author contributions

Yuzhen Dong: Conceptualisation; Formal analysis; Investigation; Methodology; Visualisation; Writing – original draft; Writing - review & editing.
Kate Nation: Conceptualisation; Funding acquisition; Investigation; Resources; Supervision; Writing – review & editing.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by a grant from the Nuffield Foundation (EDO/43392) to Kate Nation, and by resources that were made available by the Department of Children's Dictionaries and Children's Language Data at Oxford University Press, with thanks to Nilanjana Banerji and Sam Armstrong at the Oxford University Press for helpful discussions. The Oxford Children's Language Corpus is a growing database of writing for and by children developed and maintained by Oxford University Press for the purpose of children's language research.

ORCID iDs

Yuzhen Dong https://orcid.org/0000-0003-2518-9711
Kate Nation https://orcid.org/0000-0001-5048-6107

Supplemental material

Supplemental material for this article is available online.

References

Baron-Cohen, S., Golan, O., Wheelwright, S., & Granader, Y. (2010). Emotion word comprehension from 4 to 16 years old: A developmental survey. *Frontiers in Evolutionary Neuroscience*, *2*, 109. https://doi.org/10.3389/fnevo.2010.00109

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23. https://doi.org/10.1093/scan/nsw154

Beck, L., Kumschick, I. R., Eid, M., & Klann-Delius, G. (2012). Relationship between language competence and emotional competence in middle childhood. *Emotion*, *12*(3), 503–514. https://doi.org/10.1037/a0026320

Dawson, N., Hsiao, Y., Tan, A. W. M., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*, *1*(1), Article 1. https://doi.org/10.34842/5we1-yk94

Dawson, N., Hsiao, Y., Tan, A. W. M., Banerji, N., & Nation, K. (2023). Effects of target age and genre on morphological complexity in children's reading material. *Scientific Studies of Reading*, *27*(6), 529–556. https://doi.org/10.1080/10888438.2023.2206574

Dong, Y., Hsiao, Y., Dawson, N., Banerji, N., & Nation, K. (2024). The emotional content of children's writing: A data-driven approach. *Cognitive Science*, *48*(3), e13423. https://doi.org/10.1111/cogs.13423

Doost, H. T. N., Moradi, A. R., Taghavi, M. R., Yule, W., & Dalgleish, T. (1999). The development of a corpus of emotional words produced by children and adolescents. *Personality and Individual Differences*, *27*, 433–451. https://doi.org/10.1016/S0191-8869(98)00253-0

Dunn, J., Brown, J., & Beardsall, L. (1991). Family talk about feeling states and children's later understanding of others' emotions. *Developmental Psychology*, *27*(3), 448–455. https://doi.org/10.1037/0012-1649.27.3.448

J., Egbert, B., Gray & D., Biber (Eds.). (2022). Distribution considerations. In *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness* (pp. 122–155). Cambridge University Press. https://doi.org/10.1017/9781316584880.005

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Gowenlock, A. E., Norbury, C., & Rodd, J. M. (2024). Exposure to language in video and its impact on linguistic development in children aged 3–11: A scoping review. *Journal of Cognition*, *7*(1), 57. https://doi.org/10.5334/joc.385

Green, C., Keogh, K., Sun, H., & O'Brien, B. (2023). The children's picture books lexicon (CPB-Lex): A large-scale lexical database from children's picture books. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02198-y

Griffiths, S., Goh, S. K. Y., & Norbury, C. F. (2020). Early language competence, but not general cognitive ability, predicts children's recognition of emotion from facial and vocal cues. *PeerJ*, *8*, e9118. https://doi.org/10.7717/peerj.9118

Grosse, G., Streubel, B., Gunzenhauser, C., & Saalbach, H. (2021). Let's talk about emotions: The development of children's emotion vocabulary from 4 to 11 years of age. *Affective Science*, *2*(2), 150–162. https://doi.org/10.1007/s42761-021-00040-2

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology*, *55*(9), 1830–1849. https://doi.org/10.1037/dev0000686

Hogan, P. C. (2011). *What literature teaches us about emotion*. Cambridge University Press. https://doi.org/10.1017/CBO9780511976773

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

Hsiao, Y., Dawson, N. J., Banerji, N., & Nation, K. (2024). A corpus-based developmental investigation of linguistic complexity in children's writing. *Applied Corpus Linguistics*, *4*(1), 100084. https://doi.org/10.1016/j.acorp.2024.100084

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, *17*(3), 805–826. https://doi.org/10.1177/17456916211004899

Joseph, H. S. S. L., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition*, *133*(1), 238–258. https://doi.org/10.1016/j.cognition.2014.06.015

Kim, Y.-S. G., Park, C., & Park, Y. (2015). Dimensions of discourse-level oral language skills and their relation to reading comprehension and written composition: An exploratory study. *Reading and Writing*, *28*(5), 633–654. https://doi.org/10.1007/s11145-015-9542-7

Kolak, J., Monaghan, P., & Taylor, G. (2023). Language in educational apps for pre-schoolers: A comparison of grammatical constructions and psycholinguistic features in apps, books, and child-directed speech. *Journal of Child Language*, *50*(4), 895–921. https://doi.org/10.1017/S0305000922000198

Korochkina, M., Marelli, M., Brysbaert, M., & Rastle, K. (2024). The Children and Young People's Books Lexicon (CYP-LEX): A large-scale lexical database of books read by children and young people in the United Kingdom. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218241229694

Kucirkova, N. (2019). How could children's storybooks promote empathy? A conceptual framework based on developmental psychology and literary theory. *Frontiers in Psychology*, *10*, 121. https://doi.org/10.3389/fpsyg.2019.00121

Lindquist, K. A. (2017). The role of language in emotion: Existing evidence and future directions. *Current Opinion in Psychology*, *17*, 135–139. https://doi.org/10.1016/j.copsyc.2017.07.006

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs* (Vol. 1, 3rd ed., pp. xi, 366). Lawrence Erlbaum Associates Publishers.

Mar, R. A., Oatley, K., Hirsh, J., dela Paz, J., & Peterson, J. B. (2006). Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality*, *40*(5), 694–712. https://doi.org/10.1016/j.jrp.2005.08.002

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

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

Nencheva, M. L., Tamir, D. I., & Lew-Williams, C. (2023). Caregiver speech predicts the emergence of children's emotion vocabulary. *Child Development*. https://doi.org/10.1111/cdev.13897

Nook, E. C. (2021). Emotion differentiation and youth mental health: Current understanding and open questions. *Frontiers in Psychology*, *12*, 700298. https://doi.org/10.3389/fpsyg.2021.700298

Nook, E. C., Stavish, C. M., Sasse, S. F., Lambert, H. K., Mair, P., McLaughlin, K. A., & Somerville, L. H. (2020). Charting the development of emotion comprehension and abstraction from childhood to adulthood using observer-rated and linguistic measures. *Emotion*, *20*(5), 773–792. https://doi.org/10.1037/emo0000609

Ogren, M., & Sandhofer, C. M. (2021). Emotion words in early childhood: A language transcript analysis. *Cognitive Development*, *60*, 101122. https://doi.org/10.1016/j.cogdev.2021.101122

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/10.3758/s13423-014-0585-6

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science*, *21*(2), e12549. https://doi.org/10.1111/desc.12549

Schwering, S. C., Ghaffari-Nikou, N. M., Zhao, F., Niedenthal, P. M., & MacDonald, M. C. (2021). Exploring the relationship between fiction reading and emotion recognition. *Affective Science*, *2*(2), 178–186. https://doi.org/10.1007/s42761-021-00034-0

Shablack, H., & Lindquist, K. A. (2019). The role of language in emotional development. In *Handbook of emotional development* (pp. 451–478). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-17332-6_18

Siller, M., Swanson, M. R., Serlin, G., & George, A. (2014). Internal state language in the storybook narratives of children with and without autism spectrum disorder: Investigating relations to theory of mind abilities. *Research in Autism Spectrum Disorders*, *8*(5), 589–596. https://doi.org/10.1016/j.rasd.2014.02.002

Streubel, B., Gunzenhauser, C., Grosse, G., & Saalbach, H. (2020). Emotion-specific vocabulary and its contribution to emotion understanding in 4- to 9-year-old children. *Journal of Experimental Child Psychology*, *193*, 104790. https://doi.org/10.1016/j.jecp.2019.104790

Torre, J. B., & Lieberman, M. D. (2018). Putting feelings into words: Affect labeling as implicit emotion regulation. *Emotion Review*, *10*(2), 116–124. https://doi.org/10.1177/1754073917742706

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

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the Strange Stories: Revealing mentalizing impairments in autism. *Child Development*, *80*(4), 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x

Widen, S. C., & Russell, J. A. (2003). A closer look at preschoolers' freely produced labels for facial expressions. *Developmental Psychology*, *39*, 114–128. https://doi.org/10.1037/0012-1649.39.1.114

Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive Development*, *23*, 291–312. https://doi.org/10.1016/j.cogdev.2008.01.002

Wild, K., Kilgarriff, A., & Tugwell, D. (2013). The Oxford Children's Corpus: Using a children's corpus in lexicography. *International Journal of Lexicography*, *26*(2), 190–218. https://doi.org/10.1093/ijl/ecs017

Zhang, Y., & Gu, Y. (2023). A recipient design in multimodal language on TV: A comparison of child-directed and adult-directed broadcasting [Proceedings paper]. *45th Annual Conference of the Cognitive Science Society. Proceedings of the Annual Meeting of the Cognitive Science Society*, 2869–2879. Cognitive Science Society. https://escholarship.org/uc/item/17k7h7m6