Journal of Child Language (2026), 1–23
doi:10.1017/S030500092610049X

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Learning from the Input: A Corpus-Based Investigation of Chinese Classifiers in Children's Books and Child-Directed Speech

Jinyu Shi[1], Yaling Hsiao[2], Yifan Yang[1], Elizabeth Wonnacott[3] and Kate Nation[1]

[1]Department of Experimental Psychology, University of Oxford, UK; [2]School of Psychology and Clinical Language Sciences, University of Reading, UK and [3]Department of Education, University of Oxford, UK
Corresponding author: Jinyu Shi; Email: jinyu.shi@psy.ox.ac.uk

(Received 12 August 2024; revised 19 January 2026; accepted 26 January 2026)

Abstract
In Mandarin Chinese, numeral classifiers form a grammatical category that is syntactically obligatory when a noun is modified by a numeral or a demonstrative. The appropriate choice of a classifier is associated with the semantic properties of its corresponding noun and is context dependent. Experience with language is needed to learn these patterns, but little is known about how classifiers are structured in children's language environments. We compared the frequency and distribution of classifier phrases in four corpora: child-directed speech, children's television shows, children's books, and adult-directed speech. Classifier usage in children's books was more diverse than in both child-directed and adult speech. Books contained more specific classifiers that co-occurred with a higher proportion of unique nouns, whereas everyday speech relied on more generic classifiers. Books therefore provide access to classifier–noun combinations that are rare in speech. Implications for language development and language processing are discussed.

Keywords: child-directed speech; book language; Chinese classifier; corpus analysis; reading

摘要
汉语数量名结构中，量词的选择与其搭配名词的语义特征关系密切且具有语境依赖性。然而，儿童如何从语言环境中习得数量名结构，目前学界对此知之甚少。为此，本研究考察了儿童导向言语(又称儿向语)、儿童电视节目、儿童读物和成人导向言语四类语料中量词短语的使用频率与分布特征，以探讨不同类型输入对儿童量词系统习得的潜在影响。结果显示，儿童读物中量词的多样性显著高于儿童导向言语和成人导向言语，且其中低频量词的占比更高(如"股""缕"等)，这类低频量词还与数量占比更高的独特名词搭配使用;相较之下，日常言语中使用较多的则是高频通用量词(如"个""只"等)。上述结果表明，儿童读物可以为儿童提供日常言语中出现频率较低甚至完全缺失的量名搭配。这一发现凸显了儿童读物在儿童语言输入和语言发展中的独特作用。本文最后进一步讨论了阅读对语言习得与语言加工的潜在影响。

1. Introduction

Children acquire language from the input they receive. Early on, the major source of language input comes from what children hear. A large body of research has focused on the properties of child-directed speech (often in comparison with adult-directed speech) and its effect on children's language abilities (e.g. Huttenlocher et al., 2010; Jones & Rowland, 2017; Rowe, 2008; Shi et al., 2022). However, once children can read, language experience can change radically. Books written for children contain more complex language than day-to-day conversations (Cameron-Faulkner & Noble, 2013; Dawson et al., 2021; Hsiao et al., 2023; Montag, 2019; Montag et al., 2015). In turn, reading experience enhances language development (Arnold et al., 2018; Crain-Thoreson & Dale, 1992; Ece Demir-Lira et al., 2019). Much of the existing literature comes from studies of children learning English. In this paper, we expand on this by investigating the classifier–noun structure in Mandarin Chinese. Classifiers are grammatically obligatory in Chinese, yet language users have flexibility in selecting which classifier to use, making classifiers an ideal ground for studying variations in language use and exposure across modalities and a range of registers. Our focus is on the use of Chinese classifiers in three types of child-directed language input: written language in children's books, child-directed speech in daily life conversation, and media language in children's television shows and movies. As adult speakers adapt their speech for children (Cameron-Faulkner et al., 2003), we also compared child-directed language with the language used by adults in adult-directed speech to further capture the nature of classifier usage in children's books and in spoken language.

1.1. Classifiers in Mandarin Chinese

Classifiers are defined as independent morphemes or words that accompany and classify nouns based on their referent (Allan, 1977). Classifier systems are most often found in Asian languages and are rare in European languages (Her et al., 2022). In Mandarin Chinese, numeral classifiers form a grammatical category that is syntactically obligatory when a noun is modified by a numeral or a demonstrative. For example, it would be ungrammatical for Mandarin speakers to say **san mao* "three cats"; instead, the classifier 只*zhi1* is required between the numeral and the noun (三只猫*san zhi1 mao* "three CL.zhi cat").

Mandarin Chinese is rich in classifiers: the *Dictionary of Modern Chinese Classifiers Usage* (Guo, 2002) included more than 600 classifiers. As classifiers create categories for their co-occurring nouns (or the entities they refer to), they tend to be characterised by a single or multiple semantic features shared by the nouns, such as animacy, shape, dimensionality, size, and function (Allan, 1977; Gao & Malt, 2009; Saalbach & Imai, 2012). At the same time, it is not uncommon for a classifier category to contain a broad range of heterogeneous and seemingly unrelated nouns. For example, the classifier 条 *tiao2*, often proposed to classify long thin objects (Saalbach & Imai, 2012; Zhang & Schmitt, 1998), occurs with objects that fit the category, like *ropes* and *snakes*, but also *dogs*, *messages*, and *life* (see Habibi et al., 2020 on possible explanations of how classifier categories extend to different nouns). A somewhat special case is the classifier 个*ge4* (representing an individual unit). Much research has characterised 个*ge4* as the general or "default" classifier since it can precede virtually all countable nouns (Erbaugh, 2002; Gao & Malt, 2009; Guo, 2002; Zhan & Levy, 2018; but see Chen et al., 2024 for an alternative information-theoretic view on classifier defaulting). Supporting this, 个*ge4* was found to

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

be used instead of the alternative and more specific classifiers in adult speech production (Erbaugh, 1986). The choice of classifier preceding a particular noun not only differs in specificity but also varies depending on the context. For example, different classifiers can specify different referential components of a noun (e.g. the noun 课*ke* refers to a "particular lesson" when used with 堂*tang2* but an "entire module" when used with 门*men2*), or create various levels of formality (e.g. 老师 *laoshi* "teacher" used with the general classifier 个*ge4* under informal settings and 位*wei4* when expressing respect; Zhang, 2007). Given the complexity of the classifier–noun relationship, how do children acquire this system?

Several studies have explored children's comprehension of Mandarin classifiers across different ages (Chien et al., 2003; Hao et al., 2021; Li et al., 2010; Ma et al., 2023; Sera et al., 2013). For example, Li et al. (2010) asked Mandarin-speaking children to select an object or picture that could be paired with a target classifier. Results showed that while 3-year-olds (and below) performed near chance on most of the classifiers tested, accuracy increased with 4–5-year-olds. They were not only able to associate familiar objects with the target classifiers but also to generalise, that is, choose which classifier should be used with novel nouns/objects. However, these studies are based on a small set of classifiers, and the rationale for selecting particular classifiers and classifier–noun pairings is not always clear. As children learn from language input, their performance on a comprehension task will presumably vary as a function of their experience with different classifiers and classifier–noun combinations (e.g. their frequency in child-directed speech), but there is a lack of systematic investigation of this. Only Hao et al. (2021) considered the effect of classifier frequency, and their study was based on a small child-directed speech corpus, with 10 out of the 12 targeted classifier–noun pairs appearing less than 10 times in that corpus.

Turning to production, children's ability to produce appropriate classifiers for different categories of nouns emerges later than comprehension. Hao et al. (2021) used a counting task with prompts to elicit children's spoken production of classifiers for the same set of nouns/objects used in the comprehension task. They reported production accuracy (18%) to be lower than comprehension (88%). Note, however, that the majority of production "errors" used the general classifier 个 *ge4* rather than a specific classifier. While not the most optimum choice from an adult perspective, the general classifier 个 *ge4* is nevertheless grammatically correct and likely to be encountered with these nouns in some contexts. Therefore, children's apparent low production "accuracy" might reflect their reliance on more generic patterns of usage rather than using the most appropriate and specific classifier in a particular context.

Whilst these studies attempted to capture classifier usage in children's language development, they have not considered how classifiers are structured in the input that supports acquisition. Given the number and variability of classifiers in Mandarin Chinese, it is important to understand the nature of children's language experience with classifiers and how exposure to different types of language input may shape children's classifier development.

1.2. Child-directed speech and child-directed text

When talking to young children, caregivers often adapt their speech to contain simpler words, shorter utterances, and a more exaggerated prosody compared to adult-directed speech (Cameron-Faulkner et al., 2003; Soderstrom, 2007). This register is commonly

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

referred to as child-directed speech. The quality of child-directed speech seems to influence young children's language learning outcomes as more experience with diverse and sophisticated child-directed speech is associated with larger vocabulary size and more use of complex structures in children (e.g. Huttenlocher et al., 2010; Jones & Rowland, 2017; Rowe, 2008).

Alongside child-directed speech, books also form an important part of a child's language environment. Children's exposure to written language typically starts with listening to it via shared reading, followed by more independent reading once children can read. Child-directed text provides children with linguistic input that is quantitatively and qualitatively different from child-directed speech. Unlike spoken language that usually takes place in the immediate communication context, written language lacks a shared environment and cannot capitalise on extra-linguistic cues such as prosody, gestures, and facial expressions. Therefore, written texts must rely more on the text itself to convey information effectively, reflected partly in book language being more diverse and sophisticated than speech (Biber, 1991; Roland et al., 2007).

Exposure to book language allows children to experience words and syntactic structures that rarely appear in daily life speech (for review, see Nation et al., 2022). For example, Montag et al. (2015) compared word type and token frequencies in picture books targeted at young children with child-directed speech. They found that books contain more unique word types, and this difference in lexical diversity becomes bigger as the number of cumulative tokens increases. Similarly, Dawson et al. (2021) showed that picture books contain more total words, more rare words, and more structurally complex words compared to child-directed speech. There is also evidence that children's picture books contain more rare words than adult-directed speech (Massaro, 2015). Turning to sentence level, children's picture books and reading books contain a higher proportion of canonical sentences, passive sentences, and relative clauses than child-directed speech (Cameron-Faulkner et al., 2003; Hsiao et al., 2023; Montag, 2019; Montag & MacDonald, 2015), indicating greater complexity in written language. These findings suggest that the enriched linguistic input afforded by books would introduce substantial variations in children's exposure, affecting their language development (Nation et al., 2022). Previous studies have found book exposure to correlate with children's vocabulary growth (Ece Demir-Lira et al., 2019), sentence production and comprehension (Arnold et al., 2018; Montag & MacDonald, 2015), as well as academic achievement (Mol & Bus, 2011). In addition, literacy experiences predict individual differences in adult native speaker's ultimate language attainment (Dąbrowska, 2018), including their vocabulary knowledge (Chateau & Jared, 2000; Stanovich & Cunningham, 1992) and sentence processing abilities (Acheson et al., 2008; Favier & Huettig, 2021; James et al., 2018).

Another type of language input available to children is media language. Children aged 4–15 in the UK were reported to watch an average of around 18 hours of TV (including streaming) at home each week (Ofcom, 2025). It is therefore important to understand the nature and content of this language input, if we are to build a more comprehensive overview of children's language experience. Few studies have systematically examined the language exposure provided by child-directed media and its relation to child-directed speech and books. Recent corpus analyses revealed that children's books in the UK contain low-frequency and complex words not found in British television shows (including both adult- and child-directed content) (Korochkina et al., 2024); children's books also include more unique emotion words than children's television (Dong & Nation, 2025). Nevertheless, video media is popular among 3–5-year-olds, and there is some evidence that when compared to child-directed speech, it shows greater lexical

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

diversity and includes a higher proportion of rare words (Gowenlock et al., 2025). Together, these findings suggest that while media language may provide a more diverse language input than everyday speech, it does not match the complexity of book language.

At a general level, home literacy experience in Chinese correlates with children's language abilities (Liu et al., 2018; Zhang et al., 2020), but the differences between child-directed speech, child-directed media, and child-directed text have had little attention in the research literature. The variability in classifier use offers a lens through which to explore how Chinese book language differs from everyday conversation. As noted above, classifier–noun pairing in Chinese is not a one-to-one relationship and is heavily dependent on the prior discourse and extra-linguistic context. These are expected to vary across books, media, and speech since they involve different registers, content, and modalities of communication. At present, we do not know the distribution of classifiers in child-directed language, how these distributions may differ between everyday speech, media, and books, and how both of these patterns might be different from adult-directed speech. Considering the complexity of the classifier system, knowing this distribution would be an essential starting point for understanding how children experience classifiers in the language they hear or read, and how this shapes language development.

1.3. Aims

To provide a comprehensive analysis of how Mandarin classifiers are structured in children's language experience, we quantified and directly compared the use of Chinese classifiers across children's reading books and two types of spoken input: child-directed speech in daily life and in media. Additionally, we compared the three types of child-directed language with adult-to-adult speech to explore whether classifier usage varies with the target audience. With adult-directed speech being more complex than speech directed to children, this analysis allowed us to further capture how speech and written language differ.

Based on previous findings in English, we first predicted that children's books would contain a broader range and more diverse use of classifiers than children's media and child-directed speech. We also expected the general classifier 个*ge4* to be less frequent in children's books and more prominent in child-directed media and speech. Lastly, motivated by the observation that English children's picture books contain more rare words than adult-directed speech (Massaro, 2015), we predicted classifier use in adult-directed speech to be more diverse than child-directed speech, but not children's books. In our final analysis, we estimated how many classifiers and classifier–noun combinations children would encounter through varying amounts of shared book reading, and quantified how much of this input would not be available via child-directed speech alone.

2. Method

2.1. Description of corpora

We analysed four different corpora, three containing language targeted at children and one containing speech directed to adults:

2.1.1. Children's book corpus: CCLOWW

The Chinese Children's Lexicon of Written Words (CCLOWW) corpus (Li et al., 2022) includes 2131 books (both picture and reading books) for children in simplified Chinese.

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

The books are targeted at three grade levels (grades 2 and below, grades 3–4, grades 5–6, roughly mapping to ages 0–8, ages 8–10, and ages 10–12, respectively), and span fiction, nonfiction, and curriculum materials. In total, CCLOWW contains 34,672,448 character tokens.

2.1.2. Children's media corpus: CCLOOW

Data on children's language input from media came from the Chinese Children's Lexicon of Oral Words (CCLOOW) corpus (Li et al., 2023). The corpus contains 21 animated TV shows and 145 cartoon movies targeted towards 3–9-year-old children, totalling 2,745,366 character tokens.

2.1.3. Child-directed speech

These data were generated from 14 corpora in the Mandarin Chinese section of the Child Language Data Exchange System (CHILDES) database (MacWhinney, 2000). We included all available corpora except those that are transcribed in Pinyin or only had peer talk. The final set of corpora (see Supplementary Appendix A) included interactions between 1,168 children (age range from 0- to 10-years-old) and their caregivers, other family members, teachers, and researchers. The data were downloaded in CHAT format; we filtered out children's utterances using CLAN so that the dataset only included speech directed to the child. The final dataset contains 1,884,676 character tokens. Note that the contexts of the interactions included but were not limited to free play, structured play, narrative elicitation, and shared book reading. We acknowledge that the language used in shared book reading overlaps considerably with the text itself and therefore likely diverges from daily communication (Noble et al., 2018), but this overlap should only lead to a more conservative estimation of the differences between book and speech input.

2.1.4. Adult-directed speech

The adult-directed speech data were generated from two datasets. The first dataset was the Mandarin Chinese portion of the CallHome and CallFriend corpus in the conversation between adults (CABank) subsection of Talkbank (MacWhinney & Wagner, 2010). This dataset consists of 187 phone calls between Mandarin speakers from mainland China. It was downloaded in CHAT format and utterances from both interlocutors were included. The second dataset was the unscripted part of the Diversified Spoken Chinese Uttered in Social Settings (DiSCUSS) corpus (Xu et al., 2022). The dataset included 100 private dialogues (calls and in-person conversations), 80 public dialogues (lessons, broadcasts, debates, legal cross-examinations, and business transactions), and 70 monologues (e.g. spontaneous commentaries, unscripted speeches, and presentations). The two datasets generated 1,988,753 character tokens in total (582,036 from CABank; 1,406,717 from DiSCUSS).

2.2. Extraction and cleansing of classifier–noun pairings

There are two possible approaches to extracting classifiers and associated nouns: automated tools and hand coding. Given the size of our datasets, the automated method was preferred, albeit with potential accuracy trade-offs. Chinese texts pose challenges to

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

parsers due to the lack of unambiguous word boundaries like spaces and inflectional morphemes (Wong et al., 2022). This could lead to incorrect word segmentation and consequently inaccurate identification of the classifier or the noun. Fortunately, classifiers form a semi-open word class, that is, new classifiers can emerge but are limited in number and are possible to count (Erbaugh, 2006), making it feasible to manually validate the elements that were labelled as classifiers by the parsing tool. Therefore, to maximise accuracy while maintaining efficiency, we adopted automated tools for parsing and extraction, followed by manual checks on the classifiers. We established an accuracy check for the corresponding nouns and discussed the types of errors generated by our parsing and extraction method.

An overview of the final extraction process is outlined in Figure 1. We first conducted word segmentation and part-of-speech (POS) tagging using the Tsinghua University Lexical Analyzer for Chinese (THULAC, Model 2; Sun et al., 2016) implemented in Python. The parser was trained using the *People's Daily* corpus (1.2×10^7 characters) with an accuracy rate of .86 for word segmentation and .81 for POS tagging. Before parsing, we removed duplicated books and unreadable contents (rare characters, garbled characters, and non-Latin letters) as they would lead to decoding errors. After parsing, we utilised regular expressions in Python to match patterns in the POS-tagged text based on the ordering of classifiers and nouns. Classifiers were required to appear with a noun, but the noun did not have to immediately follow the classifier: cases with adjectives and other modifiers appearing in between a classifier and the corresponding noun were also included. As the parser did not provide information related to syntax, extraction was based on co-occurrences and was prone to error. Therefore, we carried out a series of modifications to refine the extraction algorithm and minimise error (see Supplementary Appendix B for the modification process and Supplementary Appendix C for the regular expressions).

To facilitate the manual validation process, we first carried out a round of data cleaning on the extracted data by filtering out those containing non-Chinese alphabets, special symbols, and Arabic numerals. Additionally, any elements tagged as classifiers that occurred only once in the entire corpus were most likely due to parsing errors, and so they were also removed. We then formed a list of unique elements that were extracted as classifiers using the parser and the algorithm, totalling 795 labelled elements. Two native Chinese speakers with backgrounds in linguistics independently hand-coded whether

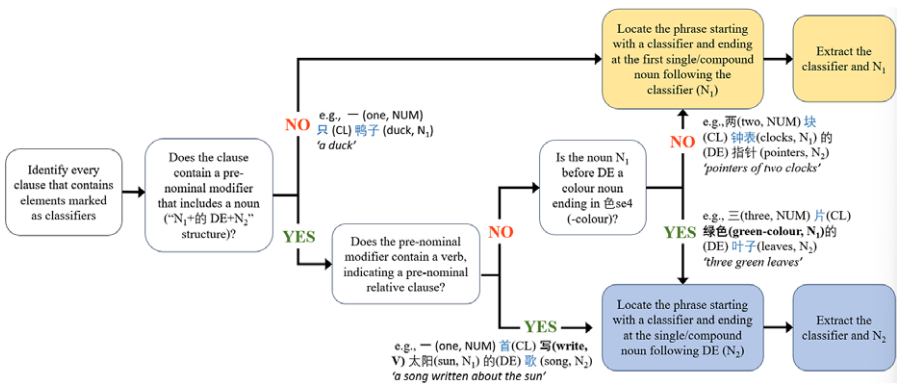

Figure 1. An overview of the extraction process with our regular expression algorithm and exemplar sentences.

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

each element can serve as a nominal classifier in Mandarin Chinese, informed by linguistic theories and corpus data. The two coders initially agreed on 91.19% of the coding (725 cases) and, after discussion, reached 100% agreement on all coding, resulting in 269 unique classifiers being identified. We then filtered the data to only include cases with these identified nominal classifiers. The removed cases consisted of verb measure words/classifiers (which serve as units for actions, e.g. *次ci4* in *看(kan,* "*saw*")*三(san,* "*three*")*次(ci4, CL)* "saw three times"), conventional measure units (e.g. *厘米limi* "centimetres"), and parsing errors.

Manual validation of classifiers can only mitigate errors to a certain extent, not to mention that it is not feasible to manually correct every individual noun. Therefore, it is crucial to establish the accuracy of the classifier–noun phrases in the dataset and understand what types of errors emerged. To do this, we randomly extracted 1000 sentences from each of the four corpora (4000 sentences in total) with the requirement that they must contain an element labelled as a noun. This avoided extracting filler utterances like "um" from the speech corpora. The first author hand-coded the nominal classifier–noun combinations that appeared in these sentences. A separate coder independently hand-coded 20% of the sentences, and the two coders reached 100% agreement on the manually coded classifier–noun pairs. We then compared the manual coding with the results generated by the automated procedure, looking at (i) the percentage of automatically identified items that are correct (i.e. precision), (ii) the percentage of manually extracted items that were correctly identified by the automated process (i.e. recall), and (iii) whether the error is related to the parser or the extraction process. We also calculated the F_1 measure, which is a weighted harmonic mean of precision and recall, operationalised as 2 × precision × recall/ (precision + recall). Table 1 lists the number of classifier–noun pairs identified in each corpus via the automated process (i.e. machine identified) and manual coding, and summarises the comparison outcomes of precision, recall, and F_1 score for classifiers, nouns, and the whole classifier–noun pairs.

There were fewer errors in child-directed speech than in the other three corpora. A possible explanation is that child-directed speech typically comprises shorter sentences with less complex pre-nominal modifications, making it easier for the parser and algorithm to correctly locate the classifiers and nouns. As expected, since classifiers were manually validated, they had a higher accuracy than nouns in three corpora. The parsing errors in classifier extraction were mostly caused by incorrect tagging of polysemous characters (e.g. *头tou2* is a classifier for domestic animals but can also be a standalone noun meaning "head"). Algorithm errors occurred when the algorithm erroneously extracted classifiers without a corresponding noun, often due to noun omission. As an example, in *每(mei, each)只(zhi1, CL)都有(douyou, all has)毒液(duye, venom)* "each one has venom," the classifier *只zhi1* serves as a referential expression for an omitted noun. Since the noun was omitted, the classifier *只zhi1* should not be extracted as part of a classifier–noun pairing, but the presence of the noun "venom" in the clause led our algorithm to incorrectly judge *只zhi1* to pair with "venom" and extract the two elements.

For nouns, parsing errors were mostly due to incorrect segmentation of multi-character words. The algorithm errors generally fell into one of two types. The first type was when the pre-nominal modification of the target noun contained another noun. For example, in *一(yi, one)位(wei4, CL)老屋(laowu, old-house)的(de, modification)邻居(linju, neighbour)* "a neighbour from the old house," the classifier *位wei4* (used with people in a polite tone) corresponds to *邻居 neighbour*, but our algorithm extracted the first noun after the classifier, leading to the incorrect extraction of *老屋 old-house* as the corresponding noun. The second error type was when two nouns were incorrectly

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

Table 1. Precision, recall, and F_1 values of classifiers, nouns, and classifier–noun pairings using automated extraction compared to manual coding in each corpus. Number of identified cases is shown in parentheses below each corpus name (machine identified/manual coding). Raw number of errors shown in parentheses below precision and recall (parsing error, algorithm error)

	Classifier			Noun			Classifier–noun pairings		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Children's books (277/296)	95.31% (7, 6)	89.19%	92.15%	90.97% (11,14)	85.14%	87.96%	86.28% (18,20)	80.74% (23,8)	83.76%
Children's media (187/188)	94.12% (4, 7)	93.62%	93.87 %	91.78% (6, 9)	91.49%	91.73%	86.10% (10, 16)	85.64% (13, 0)	85.87%
Child-directed speech (110/108)	93.64% (4, 3)	95.37%	94.50%	96.36% (2, 2)	98.15%	97.25%	90.00% (6, 5)	91.67% (5, 0)	90.83%
Adult-directed speech (287/279)	92.67% (10, 11)	95.34%	93.99%	91.98% (8, 15)	94.63%	93.29%	84.67% (18, 26)	87.10% (16, 4)	85.87%

extracted as one compound noun (e.g. in *她(ta, she) 上(shang, go to) 两(liang, two) 个(ge4, CL) 月(yue, month) 幼儿园(youeryuan, nursery) 了(le, aspect)* "she went to nursery for two months," our algorithm incorrectly extracted "month" and "nursery" as a compound noun).

3. Results

We computed the number of classifier–noun phrases, the number of unique classifiers used, and the distribution of the classifiers across the three child-directed language corpora and the adult-directed speech corpus. We expected more diverse use of classifiers and less overuse of *个ge4* in children's books compared to children's media and child-directed speech. We also predicted that adult-directed speech would contain a more diverse set of classifiers than child-directed speech, but not children's books. Having considered our questions about frequency, diversity, and distribution, we end with a simulation that estimates the total cumulative classifier and classifier–noun combinations exposure for children who engage in shared book reading at varying frequencies.

3.1. Frequency of classifier–noun phrases

First, we compared the number of classifier phrases across the four corpora. Considering the difference in corpus size, we followed Hsiao et al. (2023) and normalised the frequency of the classifiers by the total number of nouns in the corresponding corpus. In total, there were 240,886 nouns in the adult-directed corpus, 205,340 in the child-directed speech corpus, 314,640 in the child media corpus, and 5,044,476 in the children's book corpus. Figure 2a shows the frequency of classifier–noun phrases per 100 nouns in each corpus with the raw count labelled above each bar. Pairwise comparisons of proportions of classifier–noun phrases revealed that children's books contained significantly more classifier–noun pairs (7.85 classifier–noun pairs per 100 nouns) than all three speech corpora (all $p < .01$), whilst children's media contained the fewest classifier phrases (6.81 per 100 nouns, all $p < .01$). There was no significant difference in the proportion of classifier–noun phrases between child-directed speech (7.37 per 100 nouns) and adult-directed speech (7.43 per 100 nouns, $\chi^2(1) = 0.65$, $p = .42$).

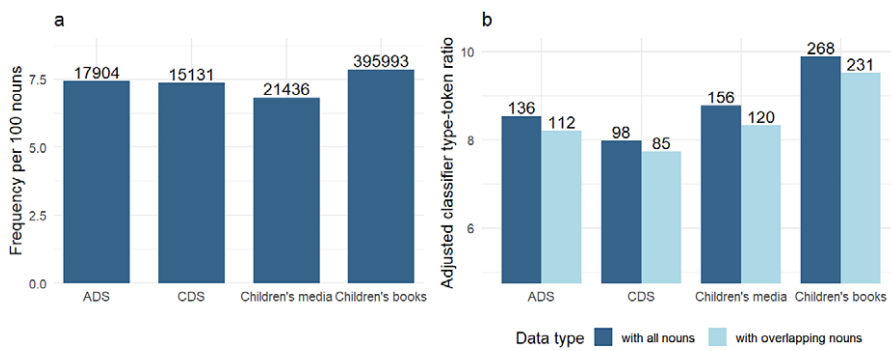

Figure 2. Frequency of classifier phrases (a) and diversity of classifiers (b) in adult-directed speech (ADS), child-directed speech (CDS), children's media, and children's books. The labels show raw count in (a) and type frequency in (b).

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

3.2. Diversity of classifiers

To compare the diversity of classifier use, we calculated the unique number of classifiers and used an adjusted type–token ratio (TTR) of classifiers for each corpus that corrects for sample size. This is less sensitive to sample size (words are more likely to be repeated when language samples contain more word tokens; Richards, 1987) than the original operationalisation as the ratio of the number of unique word types to the total number of words in a text (Templin, 1957). We chose adjusted TTR based on findings reported by Hsiao et al. (2024), who compared 24 lexical diversity measures and found that the adjusted TTR contributed most strongly to the dimension representing lexical diversity in a principal component analysis. It also showed the highest quality of representation of that dimension (as measured by cos^2) and was highly correlated with other lexical diversity measures with sampling approaches. Following Hsiao et al. (2024) and Lu (2012), the adjusted TTR was operationalised as (square of log (tokens of classifiers))/ log (tokens of classifiers/types of classifiers). Figure 2b (dark blue bars) shows that children's books contained the most diverse use of classifiers (adjusted TTR: 9.89) and the largest number of unique classifiers, followed by children's media (adjust TTR: 8.77) and adult-directed speech (adjusted TTR: 8.53), and finally by child-directed speech (adjusted TTR: 7.98). Please note that TTR calculated using a sampling-based approach produced the same overall pattern (see Supplementary Appendix D for detailed results). A linear regression model tested this trend. Within each corpus, we extracted 5% of the classifier phrases in that corpus 500 times, generating 500 random samples. We then calculated the adjusted TTR for each sample, serving as the outcome variable. The predictor variable was the corpus type and was coded using successive differences contrasts. This allowed us to test stepwise changes in classifier diversity across corpora. The results revealed a significant stepwise increase in classifier diversity from child-directed speech to adult-directed speech ($\beta = 0.72$, $SE = 0.010$, $t = 75.19$, $p < .001$), to children's media ($\beta = 0.58$, $SE = 0.010$, $t = 60.50$, $p < .001$), and to children's books ($\beta = 1.05$, $SE = 0.010$, $t = 109.79$, $p < .001$).

A potential concern is that this apparent classifier diversity might be an artefact of greater *noun* diversity in the book and media corpora: since a classifier can only be appropriately used when it is compatible with the corresponding noun, a language sample with more unique nouns provides more opportunity for different classifiers to appear. We therefore re-computed the adjusted TTR measures but limited them to classifier phrases that contained one of the 479 nouns that appeared in all four corpora. Figure 2b (light blue bars) shows that the patterns of usage remain similar to when all classifier phrases are used.

3.3. Distribution of classifiers

Having established that children's books contain a larger and more diverse range of classifiers than other registers, we next examined the distribution of classifiers. We asked whether the use of the general classifier 个 *ge4* and other high-frequency classifiers differs across corpora. Figure 3 depicts the distribution of the 30 most frequent classifiers in each corpus (the overall distribution of classifiers based on rank frequency can be found in Supplementary Appendix E), showing a Zipfian distribution for each corpus. In the figure, the *x*-axis shows the raw frequency count of each of the frequent classifiers. The labels on the right of the bars show the cumulative percentage of the observations that can be explained by each classifier and the classifiers above them. The frequency count of the

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

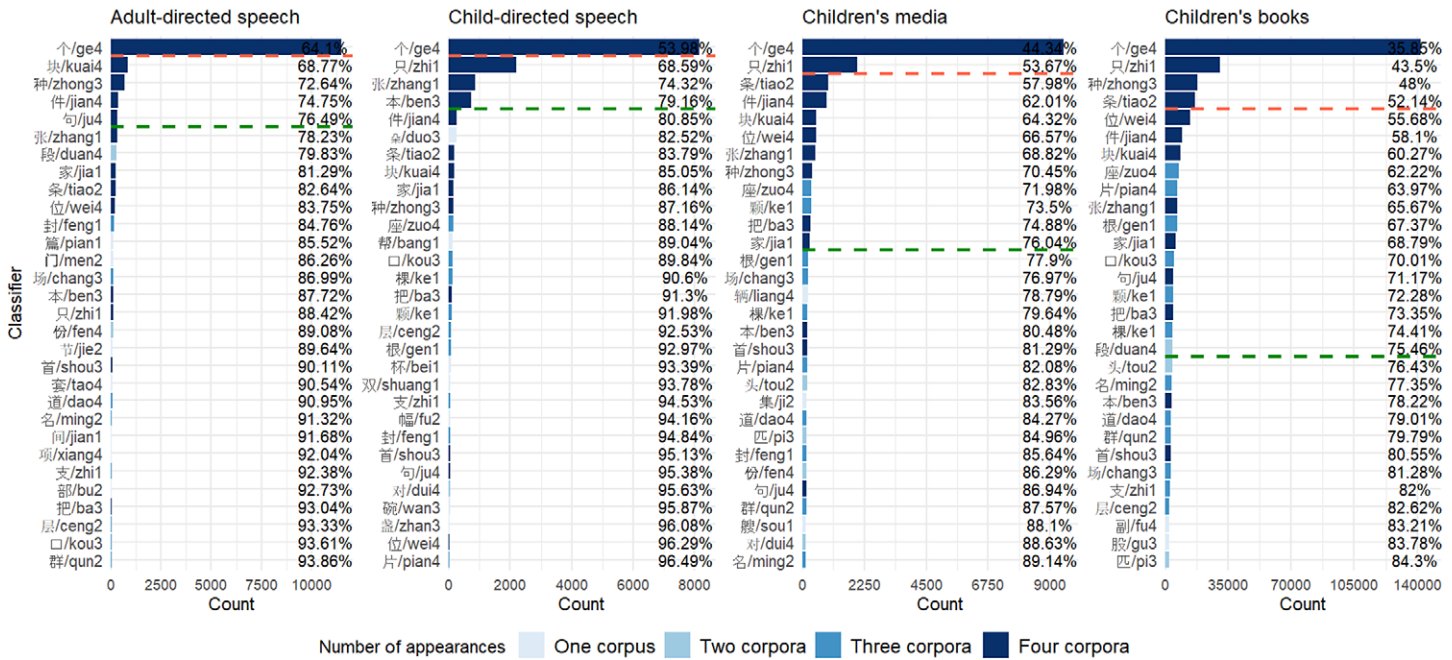

Figure 3. Frequency count of the 30 most frequent classifiers in each corpus, colour-coded based on how many of the corpora (one to four) the classifier featured within the top 30. The cumulative percentage of classifier observations out of all classifier phrases is shown in the labels on the right. The dotted lines indicate the classifiers accounted for 50% (red) or 75% (green) of classifier phrases. See text for more description.

classifiers above the red-dotted line accounts for 50% or more of all classifier phrases obtained and above the green-dotted line accounts for 75% or more of all classifier phrases. The bar colour indicates how many of the corpora (one to four) the classifier featured within the top 30.

There was a considerable overlap in highly frequent classifiers across the four corpora (indicated by the dark blue bars), especially above the first dotted line (classifiers that made up 50% or more of the data). The most common classifier in all four corpora was the general classifier 个*ge4* (an individual unit of something). In the three child-directed language corpora, 只*zhi1* (commonly used with small animals and containers) was consistently the second most common classifier. Despite these similarities, the distribution of classifiers was less spread and more clustered in both adult-directed and child-directed speech: the most common classifier 个*ge4* accounted for almost 60% of all classifier phrases. In child-directed speech, the top four most frequent classifiers made up more than 75% of the data. In contrast, the distribution of classifiers in children's books was more spread with more occurrences of the lower-frequency classifiers; children's media patterned in between child-directed speech and children's books.

Alongside these patterns in the general distribution, there were qualitative variations in the highly frequent classifiers due to the stylistic or contextual differences between the child-directed corpora. For example, the classifier 位*wei4* (used with people in a polite tone under a more formal setting) was highly frequent in children's media and children's books but ranked relatively low in child-directed speech.

Next, we looked at the distributions in terms of the unique nouns that follow the classifiers. The classifier 个*ge4* is of particular interest, as it is often proposed as the "default" classifier. We therefore examined the use of 个*ge4* with nouns in comparison with other classifiers across the four corpora, considering (1) the number of unique nouns that followed each classifier and (2) whether the nouns used with a given classifier also occurred with other classifiers. Figure 4 shows the proportion of unique nouns that co-occurred with each of the top four classifiers at least once, out of all unique nouns that appeared in classifier phrases in each corpus. Each bar is further divided into nouns that only appeared with that classifier (light blue) and nouns that also occurred with other classifiers in any of the four corpora (dark blue).

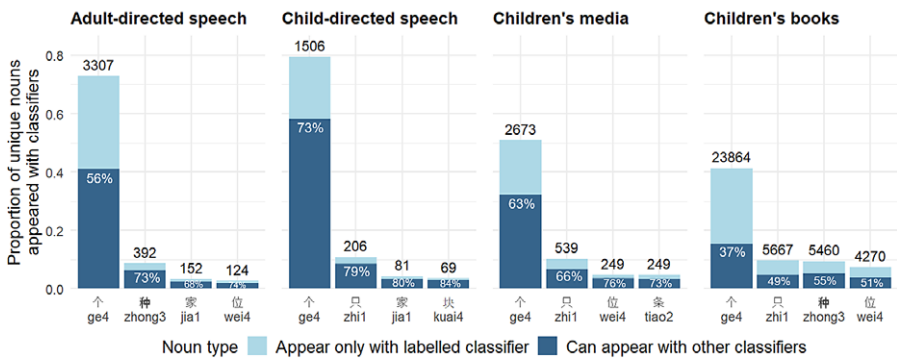

Figure 4. The top four classifiers that co-occurred with the highest number of unique noun types in each corpus. Raw counts of unique nouns are shown above the bars. Each bar is divided into nouns that only appeared with the corresponding classifier (light blue) and nouns that can occur with other classifiers (dark blue). Percentages inside the bars indicate the proportion of nouns that co-occurred with other classifiers.

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

When looking at the overall distribution, the general classifier 个 *ge4* was used with the highest proportion of unique nouns in child-directed speech (79.5% unique nouns that appeared in classifier phrases in child-directed speech had appeared with 个 *ge4* at least once), followed by adult-directed speech (73%), and by children's media (50.9%). By comparison, 个 *ge4* was used with the fewest noun types in children's books (41.3%) and instead, most unique nouns in books (58.7%) were not associated with the general classifier. Note that even in child-directed speech, around 20.5% of unique nouns that appeared in classifier phrases did not co-occur with 个 *ge4* even once. Furthermore, the category divisions within each bar show that most nouns that were used with 个 *ge4* in child-directed speech (73.2%), adult-directed speech (56.3%), and children's media (63.2%) were used with other classifiers. In children's books, however, 个 *ge4* was more often used with those nouns that do not have an alternative pairing (62.8%) compared to those that do (37.2%). This difference in 个 *ge4* usage between book and spoken language was also seen for other highly frequent classifiers.

3.4. Estimation of classifier exposure from shared book reading

Having established that children's books provide more diverse classifier input than child-directed speech and media, we next asked what this might mean in terms of actual language exposure. Specifically, we conducted a simulation to estimate the cumulative annual exposure to classifiers and classifier–noun combinations for children who engage in shared book readings at varying rates, and how much of this input would not be available through child-directed speech alone.

Following the simulation method used by Logan et al. (2019), we categorised shared book reading frequency into five levels: never (0.11 book readings per week, allowing for a very small amount of incidental reading such as one book every other month), once or twice a week (1.5 book readings per week), 4 book readings a week, 7 book readings a week, and 35 book readings a week (five books read per day). To estimate the number of classifiers and classifier–noun combinations read annually, we first calculated the median number of classifier–noun pairs per book. Since our focus is on shared book reading, we based this calculation on books targeted at Grade 2 and below (1,457 books, including 1,370 picture books). The resulting median was nine classifier–noun combinations, representing the number of classifier–noun pairs in a typical book reading session. We then estimated children's weekly exposure by multiplying this median by the number of books read per week (as defined above) and further multiplied the resulting product by 52 weeks for the yearly total. These values are reported in the "Total classifier–noun pairs" column of Table 2, shown below. Next, we estimated the number of unique classifiers and classifier–noun pairs within these totals using the adjusted TTRs generated from the book corpus targeted at Grade 2 and below (adjusted TTR for classifier = 8.78; adjusted TTR for classifier–noun pairs = 39.70). Finally, we simulated how many of these unique classifiers and pairs were present or absent in the child-directed speech. To do this, we randomly sampled a set number of classifiers and classifier–noun pairs from the Grade 2 and below book subcorpus – based on the estimated cumulative exposure at different book reading frequencies – across 500 iterations, and calculated how many items, on average, appeared in the child-directed speech dataset. The estimations are summarised in Table 2.

Based on our simulation, children who are read to daily are exposed to 3,229 more classifier–noun pairs cumulatively than their peers who are rarely or never read to. This includes an addition of 1,560 unique combinations and 106 unique classifiers. For

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

Table 2. Expected cumulative number of classifier–noun pairs and unique classifiers and classifier–noun pairs to which children are exposed annually by the number of readings per week. The number of unique classifiers and classifier–noun pairs that are present or absent in child-directed speech is shown in parentheses. All values are rounded to the closest integers. [a]Never is mathematically represented as 0.11 times per week, 1–2 is represented by 1.5 books per week, 3–5 is represented as 4 books per week, daily is represented as 7 books per week, and multiple books per day is represented as 35 books per week

Book reading frequency[a]	Total classifier–noun pairs	Unique classifier–noun combinations (present, absent)	Unique classifiers (present, absent)
Multiple books per day	16,380	5,847 (628, 5219)	155 (96, 59)
Daily	3,276	1,600 (171, 1429)	128 (80, 48)
3–5 times per week	1,872	1,006 (108, 898)	113 (70, 43)
1–2 times per week	702	439 (47, 392)	84 (52, 32)
Never	47	40 (4, 36)	22 (14, 8)

children who receive more extensive shared book reading (e.g. five books per day), the input increases further, with an addition of 13,104 classifier–noun pairs (4,247 unique classifier–noun pairs and 27 unique classifiers) than what is experienced through daily shared book reading.

4. General discussion

Focusing on Chinese classifiers, our analyses provide a comprehensive overview of the structural frequency and diversity of classifiers in Mandarin Chinese. This is the first systematic analysis of classifier usage based on large-scale language corpora. In line with our prediction, classifier use in children's books was shown to be more diverse and dispersed than both child-directed and adult-directed speech. Children's books made more use of classifiers that were more specific and far less frequent in spoken language, including in TV shows and movies. These findings extend and complement observations from English showing that children's books provide unique language input that differs from spoken language (e.g. Dawson et al., 2021; Hsiao et al., 2023; Montag, 2019).

We compared the frequency, diversity, and distribution of classifiers in adult-directed speech, child-directed speech, children's media, and children's books. The overall frequency of classifier phrases was low (around 6–8 nouns were used with classifiers per 100 nouns), with children's books containing more classifier phrases. This was not unexpected: although classifiers are grammatically required between a numeral or a demonstrative and a noun, nouns are often used in other structures. However, the low frequency of classifiers does not make investigating exposure to classifiers less relevant. On the contrary, low-frequency constructions may be particularly informative for understanding language development and how reading shapes that development and long-term attainment. Prior research has shown that individual differences in native speakers' language attainment are more pronounced for complex constructions that are rarer and more unusual than for common ones (Dąbrowska, 2018). More importantly, Dąbrowska also found that while print exposure affected linguistic knowledge in vocabulary, collocation, and grammar, the effect on vocabulary and collocation

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

was stronger because a wider range of frequencies was sampled for these two types of constructions. Therefore, the rarity of classifier–noun pairs in our analyses, especially the rare classifier–noun pairings that only appear in books, may lead to greater differences in the overall classifier input received by individuals who read at different rates. This reinforces the importance of establishing children's exposure to different classifiers across different contexts and considering this when studying classifier acquisition.

Further differences emerged in the diversity and distribution of classifiers across children's books, TV shows, and speech. Our analysis revealed a stepwise increase in the range of classifiers used from child-directed speech to adult-directed speech, children's media language, and children's books, even when controlling for noun diversity. The increased classifier diversity in adult-directed speech compared to child-directed speech is in accordance with previous observations of greater complexity in adult-directed speech (e.g. Cameron-Faulkner et al., 2003), providing further support that speakers modulate their speech based on their audience. These results are also in line with the well-documented pattern in English of words being more diverse and complex in children's books than in conversations (Dawson et al., 2021; Montag et al., 2015). Strikingly, by comparing children's books and adult-directed speech, we found that even adult–adult conversations cannot account for the range of classifiers afforded by children's books. It would be interesting to compare classifier diversity in adult books versus adult conversations. Following our findings with child-directed text and speech, and those on relative clauses in English (Montag & MacDonald, 2015), we predict classifier diversity to be greater in books written for adults.

Turning to the distribution of classifiers, as expected, in all four corpora, classifiers followed the general distribution of word frequency in language, where only a small set of words appear very frequently and the majority forms the long tail of the distribution (Dawson et al., 2021; Piantadosi, 2014; Ramscar, 2021; Zipf, 1949). This pattern is more prominent in the child-directed speech dataset, in terms of both the number of observations and the pairing of classifiers with different noun types: the four most frequent classifiers accounted for more than 75% of classifier phrases, and the most frequent classifier *个ge4* was used with around 80% of unique nouns. This distributional pattern in child-directed speech aligns with results from studies of children's own production. For example, across various paradigms, including structured play (Tse et al., 2007), story retelling (Erbaugh, 1986), and picture description (Hao et al., 2021), studies have consistently shown that children's classifier production is predominantly formed by the general classifier *个ge4.* This reflects the distribution found in spoken language input in our analyses.

Although children's books followed the same distribution, the less common and more specific classifiers accounted for a higher proportion of the classifier phrases and unique nouns in books compared to speech. In fact, despite *个ge4* being widely considered the general classifier and the default, around 58.7% of unique nouns that appeared in classifier phrases in books did not co-occur with *个ge4* even once, in comparison to the 20.5% in child-directed speech. Furthermore, amongst the nouns that co-occurred with *个ge4* in children's books, 62.80% were not used with any other alternative classifiers (both see Figure 4), meaning that *个ge4* may be their only available option.

These differences become unsurprising when we consider the discourse properties of speech versus written language. Unlike written language, speech is produced spontaneously and in the presence of a listener. From a speaker-oriented perspective, speakers have been argued to use highly frequent classifiers as they are more likely to be available

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

for spontaneous production than less frequent more specific classifiers (Zhan & Levy, 2018). Alongside the availability-based approach, research from a listener-oriented approach in other languages has proposed that speakers' use of pre-nominal modifiers like gender articles and adjectives serves to decrease the uncertainty of upcoming nouns and therefore facilitates efficient communication (Dye et al., 2018; Frantzi & Ramscar, 2022). Classifiers may serve the same purpose in Chinese. This listener-oriented explanation is supported by the finding that classifier entropy decreases with repeated mention of the noun, meaning that classifiers become less informative as a noun becomes more and more predictable in a context (Chen et al., 2024). Since spoken communication is often situated and interactive, the extra-linguistic cues provided by the broader context like the prosody of the speech, the speaker's gestures, and the visually available objects, could facilitate noun prediction, making the use of more specific (i.e. predictive) classifiers less necessary. In contrast, written language is more displaced and decontextualised, meaning that readers need to rely on the information provided by the language itself to reconstruct the meaning intended by the writer. In this view, the diverse use of classifiers in children's books is beneficial for conveying meaning in that context. It also creates rich variations in language that allow children to experience language that they rarely encounter in daily life conversations.

The language input afforded by children's TV shows and movies sits somewhat in between the language of children's books and child-directed speech. The distribution of classifiers in the media corpus was less clustered than in child-directed speech, but it still contained more occurrences of highly frequent generic classifiers compared to children's books. This makes sense as TV and movie language benefits from a shared situation and the presence of extra-linguistic cues like child-directed speech, but is not completely spontaneous, often involving well-organised scripts and preparations (Zhang & Gu, 2023). Together, these observations put media language in between day-to-day conversations and books. With media becoming a large part of children's lives, many studies have tried to tackle how screen time and video viewing relate to language development (e.g. Madigan et al., 2020; Taylor et al., 2018; see Gowenlock et al., 2024, for a scoping review), yet less is known about how language is actually structured in children's media. Our study expands on existing research and suggests that while media language is more complex than everyday conversations, it is still simpler than the language typified by books. This is consistent with recent studies comparing media language in British television to child-directed speech (Gowenlock et al., 2025) and children's books (Dong & Nation, 2025; Korochkina et al., 2024).

To get a better understanding of how the diverse classifier use in books translates to children's language experience, we also provided an estimation of the total cumulative classifier and classifier–noun bigram exposure for children who are read to at varying frequencies. The results showed that being read to daily substantially increases children's exposure to both the quantity and diversity of classifier–noun input. Amongst this additional input, a large proportion of unique classifiers and combinations are absent from everyday speech. This gap in classifier exposure further increases with more extensive shared reading. Whilst everyday speech provides exposure to the frequent classifiers and classifier–noun bigrams, book language introduces a broader range of classifiers and more fine-grained classifier–noun pairings. Our findings again highlight the important role of written language in supporting children's acquisition of a more precise classifier lexicon and classifier–noun relationship with their language experience. Based on this simulation, we predict that differences in reading experience may lead to

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

large individual differences in learning and processing of specific classifiers and classifier–noun mappings.

Furthermore, our findings also challenge the traditional approach to classifiers. Existing approaches often treat classifiers differently from other modifiers like adjectives: one of the most common assumptions for Chinese classifiers is that there is one general classifier *个ge4* that speakers default to, while the rest are considered specific (e.g. Erbaugh, 2002, 2013; Zhan & Levy, 2018). This has led previous experimental studies either to focus on the contrast between *个ge4* and other classifiers (e.g. Klein et al., 2012) or to only examine the specific classifiers as a single group (e.g. Huettig et al., 2010). However, our analyses showed that the classifier distribution is continuous and that the group of so-called specific classifiers exhibits large distributional differences. The distribution of classifiers in our study is similar to what we would predict from the distributional pattern of English word classes (Dawson et al., 2021; Montag et al., 2015), challenging the traditional binary view of classifiers. This is further supported by findings in Chen et al. (2024), where the distribution of Chinese classifiers is continuous and resembles that of Greek articles and English colour words (e.g. *个ge4* showed the same distribution as *white*, which is not considered to be the default colour adjective). Therefore, we believe that classifiers should be treated more like other pre-nominal modifiers: just like we would not say there is a default adjective in English, a single "default" item for classifiers should also not be assumed.

We recognise that our corpus of child-directed speech provides only a snapshot of the conversational contexts that children encounter. The interactions included in CHILDES typically take place indoors (e.g. at home or in a research setting) and in the context of play and other daily routines. These interactions do not capture outdoor experiences (e.g. visiting a park or going on a trip) and are constrained to particular contexts, and this may underestimate the range of classifier phrases and types of classifiers that caregivers use with their children. However, it is important to note that the differences in classifier distribution across corpora persisted even when analyses were limited to those nouns that occurred across all corpora, approximating similar contexts. Children's books still contained a wider range of classifiers than both child-directed and adult-directed speech.

More broadly, while our corpus analyses offer valuable insights into the frequency and diversity of classifiers in children's natural language environment, without experimental data, we cannot speak directly to the effects of exposure on children's language learning and how it relates to language processing. Since our findings focused on comparing the aggregated usage of classifiers across different modalities, we did not capture how classifier distribution may vary depending on the specific contexts within each modality (e.g. Chen et al., 2024 found that the distribution of classifiers differed with successive mentions of a particular noun across a text). In addition, there may also be age-related differences in classifier use, especially in child-directed speech. We conducted exploratory analyses to consider classifier diversity in speech and text directed to children of different ages and found that speakers use classifiers more often but not more diversely when conversing with older children. As for children's books, there was evidence of a developmental increase in diversity (see Supplementary Appendix F for detailed results). However, these results should be interpreted cautiously due to a lack of a large open-source dataset of Chinese child-directed speech. The limited data available for each age group do not allow us to reliably address the age-related differences. Furthermore, frequency is unlikely to be the only factor influencing the learning of classifiers and their relationship with the corresponding nouns (Ma et al., 2023). There is currently a lack of

https://doi.org/10.1017/S030500092610049X Published online by Cambridge University Press

understanding of how children gain knowledge of classifier–noun pairs, how they generalise beyond the input, and how the classifier phrases are processed as children encounter sentences. This highlights the need for experimental work that tracks individual children's language experience pre- and post-literacy and maps this to their understanding and production of classifier phrases. Longitudinal data would also provide insights into the relationship between children's experience with different types of language input and their learning and subsequent usage.

5. Conclusion

In summary, these large-scale cross-corpus analyses show that children's books provide richer and more diverse exposure to classifiers than speech and media, introducing classifier–noun pairs that are more specific and often absent from everyday conversations. Children who read (or are read to) more frequently will therefore encounter a substantially greater quantity and diversity of classifier–noun combinations, which may in turn support the development of a more precise classifier lexicon and classifier–noun mappings. Whilst our corpus-based approach provides new insights into how classifiers are structured in children's input and highlights the unique contribution of exposure to print, experimental work is needed to directly capture how these input patterns shape language acquisition and language processing.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S030500092610049X.

Data availability statement. Data and code associated with this paper are available on the Open Science Framework website (https://osf.io/tjbe3/overview).

Acknowledgements. We thank Prof. Qing Cai, Dr. Luan Li, and their team at East China Normal University who made the texts of Chinese Children's Lexicon of Written Words (CCLOWW) and Chinese Children's Lexicon of Oral Words (CCLOOW) available to us, Prof. Maosong Sun and his team for providing information on THULAC, and Jie Meng for assistance in text parsing.

Disclosure statement. ChatGPT (Version 4 and 5, published by OpenAI) was used to improve grammar and spelling and to refine code.

Financial support. The work for this paper was supported by the Clarendon Fund Scholarship awarded to Jinyu Shi and the Yinghua Scholarship to Yifan Yang.

Competing interests. The authors have no conflicts of interest to declare.

References

Acheson, D. J., **Wells, J. B.**, & **MacDonald, M. C.** (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, **40**(1), 278–289. https://doi.org/10.3758/BRM.40.1.278.

Allan, K. (1977). Classifiers. *Language*, **53**(2), 285–311.

Arnold, J. E., **Strangmann, I. M.**, **Hwang, H.**, **Zerkle, S.**, & **Nappa, R.** (2018). Linguistic experience affects pronoun interpretation. *Journal of Memory and Language*, **102**, 41–54. https://doi.org/10.1016/j.jml.2018.05.002.

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

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/10.1207/s15516709cog2706_2.

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.

Chateau, D., & **Jared, D.** (2000). Exposure to print and word recognition processes. *Memory & Cognition*, **28**(1), 143–153. https://doi.org/10.3758/BF03211582.

Chen, S., **Gibson, E.**, & **Ramscar, M.** (2024). Availability, informatively and bustiness: Why average corpus measures are an inaccurate guide to surprisal in language. In *Proceedings of the annual meeting of the cognitive science society*, **46**(46).

Chien, Y.-C., **Lust, B.**, & **Chiang, C.-P.** (2003). Chinese children's comprehension of count-classifiers and mass-classifiers. *Journal of East Asian Linguistics*, **12**(2), 91–120. https://doi.org/10.1023/A:1022401006521.

Crain-Thoreson, C., & **Dale, P. S.** (1992). Do early talkers become early readers? Linguistic precocity, preschool language, and emergent literacy. *Developmental Psychology*, **28**, 421–429. https://doi.org/10.1037/0012-1649.28.3.421.

Dąbrowska, E. (2018). Experience, aptitude and individual differences in native language ultimate attainment. *Cognition*, **178**, 222–235. https://doi.org/10.1016/j.cognition.2018.05.018.

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.

Dong, Y., & **Nation, K.** (2025). Charting the frequency and diversity of emotion words in children's language: Written language matters. *First Language*, **45**(4), 457–475. https://doi.org/10.1177/01427237251339788

Dye, M., **Milin, P.**, **Futrell, R.**, & **Ramscar, M.** (2018). Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in Cognitive Science*, **10**(1), 209–224. https://doi.org/10.1111/tops.12316.

Ece Demir-Lira, Ö., **Applebaum, L. R.**, **Goldin-Meadow, S.**, & **Levine, S. C.** (2019). Parents' early book reading to children: Relation to children's later language and literacy outcomes controlling for other parent language input. *Developmental Science*, **22**(3), e12764. https://doi.org/10.1111/desc.12764.

Erbaugh, M. S. (1986). Taking stock: The development of Chinese noun classifiers historically and in young children. In C. Craig (Ed.), *Noun Classes and Categorization*, (Vol. **7**, pp. 399–436). John Benjamins Publishing Company.

Erbaugh, M. S. (2002). Classifiers are for specification: Complementary functions for Sortal and general classifiers in Cantonese and mandarin. *Cahiers de Linguistique Asie Orientale*, **31**(1), 33–69. https://doi.org/10.1163/19606028-90000098.

Erbaugh, M. S. (2006). Chinese classifiers: Their use and acquisition. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The handbook of east Asian psycholinguistics* (1st ed., pp. 39–51). Cambridge University Press. https://doi.org/10.1017/CBO9780511550751.005.

Erbaugh, M. S. (2013). Classifier choices in discourse across the seven main Chinese dialects. In Z. Jing-Schmidt (Ed.), *Studies in Chinese Language and Discourse* (Vol. **2**, pp. 101–126). John Benjamins Publishing Company; Portico. https://doi.org/10.1075/scld.2.05erb

Favier, S., & **Huettig, F.** (2021). Long-term written language experience affects grammaticality judgements and usage but not priming of spoken sentences. *Quarterly Journal of Experimental Psychology*, **74**(8), 1378–1395. https://doi.org/10.1177/17470218211005228.

Frantzi, I., & **Ramscar, M.** (2022). *The structure of Greek gender classes and how they smooth signalling in noun phrases*. [Bachelor Thesis]. Universität Tübingen.

Gao, M. Y., & **Malt, B. C.** (2009). Mental representation and cognitive consequences of Chinese individual classifiers. *Language and Cognitive Processes*, **24**(7–8), 1124–1179. https://doi.org/10.1080/01690960802018323.

Gowenlock, A., **Rodd, J.**, **Malory, B.**, & **Norbury, C.** (2025). *A comparison of lexical features in children's video media and child-directed speech*. EPS Lancaster meeting.

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.

Guo X. (2002). *现代汉语量词用法词典 [Dictionary of modern Chinese classifiers usage]*. Language & Culture Press.

Habibi, A. A., **Kemp, C.**, & **Xu, Y.** (2020). Chaining and the growth of linguistic categories. *Cognition*, **202**, 104323. https://doi.org/10.1016/j.cognition.2020.104323.

Hao, Y., **Bedore, L.**, **Sheng, L.**, **Zhou, P.**, & **Zheng, L.** (2021). Exploring influential factors of shape classifier comprehension and production in mandarin-speaking children. *First Language*, **41**(5), 573–604. https://doi.org/10.1177/01427237211026435.

Her, O.-S., **Hammarström, H.**, & **Allassonnière-Tang, M.** (2022). Defining numeral classifiers and identifying classifier languages of the world. *Linguistics Vanguard*, **8**(1), 151–164. https://doi.org/10.1515/lingvan-2022-0006.

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.

Huettig, F., **Chen, J.**, **Bowerman, M.**, & **Majid, A.** (2010). Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. *Journal of Cognition and Culture*, **10**(1–2), 39–58. https://doi.org/10.1163/156853710X497167.

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.

James, A. N., **Fraundorf, S. H.**, **Lee, E.-K.**, & **Watson, D. G.** (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, **102**, 155–181. https://doi.org/10.1016/j.jml.2018.05.006.

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.

Klein, N. M., **Carlson, G. N.**, **Li, R.**, **Jaeger, T. F.**, & **Tanenhaus, M. K.** (2012). Classifying and massifying incrementally in Chinese language comprehension. In D. Massam (Ed.), *Count and mass across languages*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654277.003.0014.

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*, 17470218241229694. https://doi.org/10.1177/17470218241229694.

Li, L., **Yang, Y.**, **Song, M.**, **Fang, S.**, **Zhang, M.**, **Chen, Q.**, & **Cai, Q.** (2022). CCLOWW: A grade-level Chinese children's lexicon of written words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01890-9.

Li, L., **Zhao, W.**, **Song, M.**, **Wang, J.**, & **Cai, Q.** (2023). CCLOOW: Chinese children's lexicon of oral words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02077-6.

Li, P., **Huang, B.**, & **Hsiao, Y.** (2010). Learning that classifiers count: Mandarin-speaking children's acquisition of sortal and mensural classifiers. *Journal of East Asian Linguistics*, **19**(3), 207–230. https://doi.org/10.1007/s10831-010-9060-1.

Liu, C., **Georgiou, G. K.**, & **Manolitsis, G.** (2018). Modeling the relationships of parents' expectations, family's SES, and home literacy environment with emergent literacy skills and word reading in Chinese. *Early Childhood Research Quarterly*, **43**, 1–10. https://doi.org/10.1016/j.ecresq.2017.11.001.

Logan, J. A. R., **Justice, L. M.**, **Yumuş, M.**, & **Chaparro-Moreno, L. J.** (2019). When children are not read to at home: The million word gap. *Journal of Developmental and Behavioral Pediatrics*, **40**(5), 383–386. https://doi.org/10.1097/dbp.0000000000000657.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, **96**, 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x.

Ma, W., **Zhou, P.**, & **Golinkoff, R. M.** (2023). The role classifiers play in selecting the referent of a word. *Language*, **8**(1). https://doi.org/10.3390/languages8010084.

MacWhinney, B. (2000). *The Childes project: Tools for Analyzing talk. Transcription format and programs*. Psychology Press.

MacWhinney, B., & **Wagner, J.** (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung: Online-Zeitschrift Zur Verbalen Interaktion*, **11**, 154–173.

Madigan, S., **McArthur, B. A.**, **Anhorn, C.**, **Eirich, R.**, & **Christakis, D. A.** (2020). Associations between screen use and child language skills: A systematic review and meta-analysis. *JAMA Pediatrics*, **174**(7), 665–675. https://doi.org/10.1001/jamapediatrics.2020.0327.

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.

Mol, S. E., & **Bus, A. G.** (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, **137**, 267–296. https://doi.org/10.1037/a0021890.

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., & **MacDonald, M. C.** (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, **144**(2), 447.

Nation, K., **Dawson, N. J.**, & **Hsiao, Y.** (2022). 'Book language and its implications for children's language, literacy, and development': Corrigendum. *Current Directions in Psychological Science*, **31**, 464–464. https://doi.org/10.1177/09637214221119448.

Noble, C. H., **Cameron-Faulkner, T.**, & **Lieven, E.** (2018). Keeping it simple: The grammatical properties of shared book reading. *Journal of Child Language*, **45**(3), 753–766. https://doi.org/10.1017/S0305000917000447.

OFCOM. (2025). *Children and parents: Media use and attitudes report 2025*. OFCOM. https://www.ofcom.org.uk/media-use-and-attitudes/media-habits-children/children-and-parents-media-use-and-attitudes-report-2025

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.

Ramscar, M. (2021). How children learn to communicate discriminatively. *Journal of Child Language*, **48**(5), 984–1022. https://doi.org/10.1017/S0305000921000544.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language*, **14**(2), 201–209. https://doi.org/10.1017/S0305000900012885.

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.

Saalbach, H., & **Imai, M.** (2012). The relation between linguistic categories and cognition: The case of numeral classifiers. *Language and Cognitive Processes*, **27**(3), 381–428. https://doi.org/10.1080/01690965.2010.546585.

Sera, M. D., **Johnson, K. R.**, & **Kuo, J. Y.** (2013). Classifiers augment and maintain shape-based categorization in mandarin speakers. *Language and Cognition*, **5**(1), 1–23. https://doi.org/10.1515/langcog-2013-0001.

Shi, J., **Gu, Y.**, & **Vigliocco, G.** (2022). Prosodic modulations in child-directed language and their impact on word learning. *Developmental Science*. https://doi.org/10.1111/desc.13357.

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.

Stanovich, K. E., & **Cunningham, A. E.** (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition*, **20**(1), 51–68. https://doi.org/10.3758/BF03208254.

Sun, M., **Chen, X.**, **Zhang, K.**, **Guo, Z.**, & **Liu, Z.** (2016). *Thulac: An efficient lexical analyzer for Chinese*. [Computer software]. Tsinghua University. https://thulac.thunlp.org/

Taylor, G., **Monaghan, P.**, & **Westermann, G.** (2018). Investigating the association between children's screen media exposure and vocabulary size in the UK. *Journal of Children and Media*, **12**(1), 51–65. https://doi.org/10.1080/17482798.2017.1365737.

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships*. University of Minnesota Press.

Tse, S. K., **Li, H.**, & **Leung, S. O.** (2007). The acquisition of Cantonese classifiers by preschool children in Hong Kong. *Journal of Child Language*, **34**(3), 495–517. https://doi.org/10.1017/S0305000906007975.

Wong, K.-F., **Li, W.**, **Xu, R.**, & **Zhang, Z.** (2022). *Introduction to Chinese natural language processing*. Springer Nature.

Xu, J., **Dong, T.**, **Sun, M.**, **Chen, Z.**, **Liu, F.**, **Wang, B.**, **Wang, Y.**, **Li, Y.**, **Wang, Y.**, **Ma, B.**, **Liu, Z.**, **Qian, Y.**, **Zhu, Z.**, **Quan, L.**, & **Lu, J.** (2022). *The BFSU DiSCUSS corpus: The corpus of Diversified Spoken Chinese Uttered in Social Settings (DiSCUSS)* [data set]. Beijing Foreign Studies University. https://corpus.bfsu.edu.cn/info/1070/1335.htm

Zhan, M., & **Levy, R. P.** (2018). Comparing Theories of speaker choice using a model of classifier production in Mandarin Chinese. Prof. Levy. https://doi.org/10.18653/v1/n18-1181

Zhang, H. (2007). Numeral classifiers in mandarin Chinese. *Journal of East Asian Linguistics*, **16**(1), 43–59. https://doi.org/10.1007/s10831-006-9006-9.

Zhang, S., & **Schmitt, B.** (1998). Language-dependent classification: The mental representation of classifiers in cognition, memory, and ad evaluations. *Journal of Experimental Psychology: Applied*, **4**(4), 375–385. https://doi.org/10.1037/1076-898X.4.4.375.

Zhang, S.-Z., **Inoue, T.**, **Shu, H.**, & **Georgiou, G. K.** (2020). How does home literacy environment influence reading comprehension in Chinese? Evidence from a 3-year longitudinal study. *Reading and Writing*, **33**(7), 1745–1767. https://doi.org/10.1007/s11145-019-09991-2.

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

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

Cite this article: Shi, J., Hsiao, Y., Yang, Y., Wonnacott, E., & Nation, K. (2026). Learning from the Input: A Corpus-Based Investigation of Chinese Classifiers in Children's Books and Child-Directed Speech. *Journal of Child Language* 1–23, https://doi.org/10.1017/S030500092610049X