



CCLOOW: Chinese children's lexicon of oral words

Luan Li^{1,2,3} · Wentao Zhao^{1,2,3} · Ming Song^{1,2,3} · Jing Wang^{1,2,3} · Qing Cai^{1,2,3}

Accepted: 16 January 2023
© The Psychonomic Society, Inc. 2023

Abstract

In this article, we introduce the Chinese Children's Lexicon of Oral Words (CCLOOW), the first lexical database based on animated movies and TV series for 3-to-9-year-old Chinese children. The database computes from 2.7 million character tokens and 1.8 million word tokens. It contains 3920 unique character and 22,229 word types. CCLOOW reports frequency and contextual diversity metrics of the characters and words, as well as length and syntactic categories of the words. CCLOOW frequency and contextual diversity measures correlated well with other Chinese lexical databases, particularly well with that computed from children's books. The predictive validity of CCLOOW measures were confirmed with Grade 2 children's naming and lexical decision experiments. Further, we found that CCLOOW frequencies could explain a considerable proportion in adults' written word recognition, indicating that early language experience might have lasting impacts on the mature lexicon. CCLOOW provides validated frequency and contextual diversity estimates that complements current children's lexical database based on written language samples. It is freely accessible online at <https://www.learn2read.cn/ccloow>.

Keywords Chinese · Children · Lexical database · Frequency · Movies and TV series

Introduction

Word frequency (WF) is arguably the most established predictor of lexical processing accuracy and efficiency (for a review, see Brysbaert et al., 2018). Obtaining reliable frequency estimates has thus become one of the most pursued aims in psycholinguistic research in the past two decades. Intriguingly, the extent to which WF explains word recognition might depend on the language register from which the frequency values were computed (Brysbaert & New, 2009; Brysbaert et al., 2018). The best frequency measures

for undergraduate students are based on corpora of television and movie subtitles (Brysbaert & New, 2009) or social media (Gimenes & New, 2016). Whereas traditional frequency measures based on printed books and newspapers seem to better account for the language experience of the elderly (Brysbaert & Ellis, 2016). What about children? Particularly, what are good frequency measures for beginner readers who are starting to get exposed to printed books while there is also considerable media language input? In this article, we present a Chinese lexical database based on animated movies and TV series for young children: Chinese Children's Lexicon of Oral Words (CCLOOW). We show that CCLOOW word frequencies, as well as contextual diversity measures, can explain young children's word recognition performance while frequency measures based on adults' language samples cannot. They could also explain unique variance in adults' word recognition latencies in addition to adult frequency measures.

How to compute WF measures

The necessity of WF norms comes from the robust findings that high-frequency words enjoy a processing advantage relative to low-frequency words. The frequency effect has been consistently demonstrated in lexical processing tasks such as

✉ Jing Wang
wangjing@psy.ecnu.edu.cn

✉ Qing Cai
qcai@psy.ecnu.edu.cn

¹ Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), Affiliated Mental Health Center (ECNU), Institute of Brain and Education Innovation, School of Psychology and Cognitive Science, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, China

² Shanghai Changning Mental Health Center, Shanghai, China

³ Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

lexical decision (Balota et al., 2007; Tsang et al., 2018; Van Heuven et al., 2014), word naming (Balota et al., 2004; Liu et al., 2007) and picture naming (Lampe et al., 2021; Taylor et al., 2012). Particularly, WF is the most important variable in predicting lexical decision reaction times (RTs), accounting for over 40% of the variance (Balota et al., 2007). In comparison, when frequency is partialled out, the contribution of other lexical variables (e.g., familiarity, imageability, family size etc.) together contribute less than 5% of the variance in RT (Balota et al., 2004; Yap et al., 2009). Given the importance of WF, it is not surprising that psycholinguistic researchers have spent considerable efforts in seeking reliable frequency measures.

The type of corpora based on which WF is calculated has gone through some changes. WF is traditionally computed by counting words' occurrences in written samples of newspapers, magazines, and scientific books. An example is the widely used Kucera and Francis' (Kucera & Francis, 1967) word list. With the increasing popularity of Internet use, web-based frequency measures have also been created (Balota et al., 2004; Burgess & Livesay, 1998; Herdağdelen & Marelli, 2017). Later in Brysbaert and New's (2009) seminal work, television and movies subtitles were proposed to be a good estimate of the language experience of average adult language users. Importantly, as megastudies of word recognition data have become available, the predictive validity of traditional norms such as KF frequency on lexical decision and word naming was called into question (Balota et al., 2004, 2007). Compared with frequency measures based on subtitles, the KF norms explained 6% and 10% less variance in word recognition RTs and accuracies respectively (Brysbaert & New, 2009). It was suggested that movie subtitles are reflective of the language the participants are most likely exposed to, whereas traditional frequency measures were based on very limited topics and edited language that is detached from natural language use. Because of their high predictive validity, subtitle-based frequency norms quickly became available in many different languages such as Chinese (Cai & Brysbaert, 2010), German (Brysbaert et al., 2011), Dutch (Keuleers et al., 2010) and Spanish (Cuetos et al., 2011). They also showed better performance in explaining word recognition than traditional frequency norms. For example, the SUBTLEX-CH frequencies (Cai & Brysbaert, 2010) contributed 42.7% of the variance in adults' lexical decision latencies, compared with 13.7–27.3% of the variance explained by frequencies based on written language samples.

Another important issue is the computation of standardized frequencies. Frequency measures have mostly been raw counts or frequency per million words (fpmw). However, Van Heuven et al. (2014) noted that the interpretation of these measures depends highly on the size of the corpus. Frequency values in a 1-million-word corpus are not

comparable to that in a 20-million-word corpus, because the smallest value was 1 fpmw in the former while the frequency could be less than 1 fpmw in the latter, despite both being the lowest in respective corpora. Therefore, they proposed a frequency measure that promotes cross-corpora interpretation: the *Zipf* scale. It is a logarithmic scale roughly ranging from 1 to 7, computed as: $\log_{10}(\text{frequency per million words}) + 3$. A Zipf value smaller than 1 (corresponding to one occurrence per 100 million words) represents very low frequency; a value between 3 (corresponding to one occurrence per million words) and 4 indicates medium frequency, and a value greater than 6 (corresponding to 1 occurrence per 1000 words) suggests very high frequency.

More recently, a complementary or alternative index to WF has been proposed to take into account the contextual variations in which words are experienced – contextual diversity (CD). It is calculated by counting the number of unique documents in which a word occurs (Adelman et al., 2006) such that within-document redundant appearances are accounted for. It has been argued that CD might outperform WF in reflecting the linguistic experience with a word. According to the Semantic Distinctiveness Model (Johns & Jones, 2022; M. Jones et al., 2012), the strength of a word's lexical representation is updated each time it is encountered in a new context. The update is determined by the dissimilarity between the current context and the previous contexts the word occurs in. Thus, words experienced in a wider variety of contexts, having a high CD, should be more easily acquired and processed than low-CD words, as they are more likely to be accessed in the future. There is indeed evidence that CD explains more variance in word recognition latencies than WF (Adelman et al., 2006; Brysbaert & New, 2009). In sentence reading, high-CD words were fixated for a shorter time than low-CD words, and once CD was controlled for, the WF effect was found diminished (Chen, Huang, et al., 2017a; Chen, Zhao, et al., 2017b; Plummer et al., 2014). The CD effect has also been lately demonstrated in children (Huang et al., 2020; Perea et al., 2013). More diverse contexts also seem to facilitate learning to read novel words (Joseph & Nation, 2017; Pagan & Nation, 2019; Rosa et al., 2017).

Lexical databases for children

Turning to developmental research, the number of frequency databases based on children's reading materials has been growing rapidly in recent years. Some of the earliest children's databases were for English such as the Children's Printed Word Database (CPWD, Masterson et al., 2003) and its later extension (Masterson et al., 2010). CPWD was compiled from 556 books used by Grades 1–4 teachers in a representative sample of elementary schools in UK. It provides WF as well as orthographic and phonological variables

for 12,193 English words for 5-9-year-old children. Children's lexical databases in other languages have followed ever since (French: L  t   et al., 2004; Chinese: Li et al., 2022; German: Schroeder et al., 2014; Greek: Terzopoulos et al., 2017). Most of these corpora sampled from grade-leveled school textbooks (e.g., Terzopoulos et al., 2017) with the aim to reflect a developmental trend of average children's experience with printed words. Some of them also included measures of CD, calculated as the proportion of textbooks in which a word occurs at a given grade level (Soares et al., 2014; Terzopoulos et al., 2017).

Constructing specialized lexical corpora for children is necessitated by the drastically different language environments children and adults are exposed to. Infants and toddlers mostly hear child-directed speech from caregivers, which are characterized by item-based phrases (Cameron-Faulkner et al., 2003) and decontextualized language such as narratives (Rowe, 2012). When children start to read, the picture books they most likely read are accompanied by illustrative pictures and use more communicative and interactive language than adults' printed books. Analyses have shown that word frequencies based on children's materials correlated with each other better than with adults' frequency norms (Van Heuven et al., 2014). Given that the best frequency measures are based on the language samples the participants are most familiar with (Brysbart et al., 2018), there is good reason to predict that frequencies computed from children's materials are better predictors of children's lexical processing than adults' frequencies.

In Chinese, there are currently two lexical databases that provide WF and CD measures based on children's reading materials. The CJC database (Huang et al., 2020) is based on a collection of 52 Grades 1–4 textbooks used in Jiangsu Province and 43 storybooks, containing 2.65 million characters and 1.83 million words. Using the CJC WF and CD measures, the researchers have found that Grade 4 children's RTs in lexical decision were affected by CD and not WF. The other database is the recently released Chinese Children's Lexicon of Written Words (CCLOWW, Li et al., 2022), which is by far the largest written lexical database for Chinese children and the only one publicly available. CCLOWW sampled 2131 books of 34 million characters (22 million words) organized into three grade levels: Grades 2 and below (G2), Grades 3–4 (G34) and Grades 5–6 (G56). In particular, the G2 subcorpus was compiled from over 1500 picture books and an additional couple of Grades 1–2 textbooks. Thus, they should reflect the very early print experience of preschoolers and beginner readers. CCLOWW frequencies were not only able to explain a fair amount of variance in Grade 3 children's word naming RTs, but also, they contributed significant extra variance in adults' naming and lexical decision RTs in addition to adult frequency measures, indicating the influence of early reading experience on the mature lexicon and lexical processing.

Subtitle-like frequencies for children?

Despite the popularity of adult frequency norms based on movie subtitles, there is a lack of children's lexical database in similar language registers. SUBTLEX-UK (Van Heuven et al., 2014) is the only corpus that contained subtitles from two children's channels broadcasted in UK: CBeebies for 0-6-year-old children and CBBC for 6-12-year-old children. The subtitle-based frequencies were correlated with the children's CPWD written frequencies. It was found that CPWD correlated with CBeebies and CBBC frequencies (CBeebies: $r = .756$; CBBC: $r = .690$) better than with SUBTLEX-UK frequencies ($r = .664$). Intriguingly, although CBeebies frequencies were assumed to be based on subtitles of TV channels, we know that when preschool children watch TV, they cannot yet read the subtitles, if there are any. This is, for prereaders, they are more likely a source of spoken rather than written language input.

The impact of screen media on children's language development is two-fold. Although watching TV and using other screen media have sometimes been negatively associated with language development because it may displace other language-learning activities (Pagani et al., 2013), it can also be beneficial provided that it delivers interactive content (Linebarger & Vaala, 2010; Myers et al., 2017) and the time spent on the screen was not excessive (Dore et al., 2020). Some media content, such as animated cartoons and movies, could potentially provide a learning environment that promotes interactive learning similar to daily speech, because the language is also accompanied by communicative cues from the characters. Processing such multisensory information might enhance semantic memory retention (Li & Jeong, 2020; Mayer et al., 1999) and may thus lead to a learning advantage relative to unimodal learning environments such as reading printed texts.

Importantly, screen media such as television does constitute a considerable portion of children's everyday language exposure. Recent surveys show that on average, children under the age of three in the United States spend as much as 3.6 h a day watching TV (Madigan et al., 2019). It is likely that beginner readers spend comparable or more time on screen than they read books. Therefore, WF computed from children's TV programs and movies might provide a good reflection of their daily language experience.

The CCLOWW database

In this article, we introduce the Chinese Children's Lexicon of Oral Words (CCLOWW), a new database of characters and words compiled from animated movies and TV series for 3-to-9-year-old Chinese children. We aim to provide an addition to the recent CCLOWW database to profile the lexical experience of prereaders and beginner readers

in mainland China. Importantly, with a series of validation analyses, we show that frequency norms for children could be reliably computed from screen media language similar to subtitle-based databases for adults (e.g., Cai & Brysbaert, 2010).

In the following, we first describe how the corpus was collected and how the language samples were processed. We then provide information about the distributions of word length, syntactic category and then frequency and CD indices of the characters and words in CCLOOW. Next, we evaluated the predictive validity of CCLOOW character frequencies and CDs first with existing character naming data of Grades 2–3 children. We then conducted word naming and lexical decision experiments with Grade 2 children to validate the word measures. We also compare the effects of CCLOOW word frequencies with that of CLOWW G2 subcorpus and SUBTLEX-CH frequencies on the beginner readers' lexical recognition performance. This way, we examine the notion that the best WF norms are based on language to which the participants are most likely exposed. Finally, we investigated the effects of CCLOOW word frequencies on adults' word recognition latencies to explore the potential influences of early language experience on the mature lexicon.

The CCLOOW database

Corpus sampling and linguistic processing

We acquired a sample of 21 animated TV series and 145 animated movies in Mandarin Chinese. To ensure that the materials were representative of what young Chinese children commonly watch, we checked with the “child” sections of popular streaming platforms (Tencent Video, iQiyi, Youku) and the China Central Television children's channel to make sure that the samples were included in these programs. Note that there is no official recommended age for the TV series and movies collected for the current corpus. Our aim was to collect a sample that likely covers average 3-to-9-year-old children's exposure to screen media language in mainland China. The samples were all narratives. The audio materials were converted to written samples using the AISPEECH recognition software (<https://www.aispeech.com/>, Gong et al., 2022; Liu et al., 2020). The written samples were then cleaned and proofread by three native Chinese speakers. Each movie and episode of TV series was treated as one document. To avoid overrepresentations of any TV series with multiple episodes when calculating CD, 20 of the series that contained more than ten episodes and an average episode length of less than 300 characters were divided into documents of approximately equal length comprising

multiple episodes. The final corpus thus contained a total of 265 documents.

Word segmentation was conducted using fastHan (Geng et al., 2020), a BERT (Bidirectional Encoder Representations from Transformers)-based Chinese natural language processing toolkit. FastHan has been tested with multiple corpora and has achieved 97.41% accuracy on word segmentation and 95.66% on POS tagging, outperforming other popular Chinese word segmentation tools (Geng et al., 2020). It used the Penn Chinese Treebank 9.0 (Xue et al., 2019) for POS-tagging and dependency parsing. The output of fastHan was cleaned following Li et al. (2022). This is, we first removed non-characters including alphabet letters and numbers in low-frequency sequences. The Table of General Standard Chinese Characters (Ministry of Education, 2013) was then used to exclude characters that are no longer used in contemporary Chinese. Next, the following items were removed: items with more than 15 characters and a frequency count smaller than 2, non-nouns with more than eight characters, interjections with more than four characters and numerals with a frequency count smaller than 10. These are likely nonwords.

Frequency and contextual diversity calculation

We calculated the total number of times a character/word occurs in the corpus (raw frequency), the number of occurrences per million characters/words, and the number of unique documents in which a character/word occurs (CD). In addition, we computed the standardized frequency measure Zipf (van Heuven et al., 2014) using the formula: $\text{Zipf} = \log_{10}(\text{frequency per million}) + 3$. We also computed log-transformed values for contextual diversity as $\log_{10}(\text{CD})$. In the following, we report distribution and experimental analysis using Zipf and $\log_{10}(\text{CD})$ values. Other indices can be found and downloaded in the online database.

Summary of the corpus

The final corpus comprised 2,745,366 character tokens and 1,889,656 word tokens. There were 3920 unique character types and 22,229 unique word forms (26,582 non-lemmatized words). The mean document length was 10,360 characters (min = 342, max = 60,709) and 7,131 words (min = 189, max = 34,394). Compared with CLOWW G2 (2.2 million character and 1.5 million word tokens), which contained 4351 character and 37,516 word types, CCLOOW has much fewer unique items, despite its larger size. This shows that written texts for young Chinese children provide a wider range of lexical items than the movies and TV series they watch, which is broadly consistent with Montag et al.'s (2015) finding that children's picture books (written texts) contain more diverse word types than child-directed speech (spoken language).

Statistics of the characters and words

Word length Among the words in CCLOOW, 1891 contained one character, 15,636 two characters, 3453 three characters and 1154 four characters. Two-character words were the most common type of words, accounting for over 70% of all word types, which is consistent with previous findings with Chinese dictionary (Tan & Perfetti, 1999) and children’s corpora (L. Li et al., 2022). Nevertheless, one-character words were most frequent (1,889,656 tokens), accounting for over 60% of all word tokens. This result adds to previous findings that one-character words take up 58.41% of the CCLOWW Grade 2 and below (G2) subcorpus and the percentage decreases to 52.94% in the Grades 5–6 (G56) subcorpus. Together, they show that children encounter more complex words as they develop from prereaders (CCLOOW, CCLOWW G2) to advanced readers (CCLOWW G56).

Syntactic categories Table 1 presents the percentages of word types and tokens by syntactic categories in CCLOOW. The distributions in CCLOWW G2 are also shown to allow direct comparisons. The percentage ranks of words’ syntactic categories were largely consistent with CCLOWW G2. Also, similar to CCLOWW G2, content words (nouns, verbs, adjectives, adverbs, numerals, and measure words) take up over 95% of all word types. Nouns were the most typical type of words, followed by verbs, adjectives, and adverbs. Although over 50% of the word types are nouns, verb tokens accounted for a much higher percentage than noun tokens. This pattern is also demonstrated in CCLOWW G2 but to a lesser degree. Another interesting comparison between CCLOOW and CCLOWW G2 is the percentages of some function words. For example, pronoun types took up similar proportions in the two databases (.35 in CCLOOW and .29 in CCLOWW G2), whereas its token percentage was much higher in CCLOOW (12.56%) than in CCLOWW G2 (7.95%). This potentially reflects that media language afford a communicative context where some cues are present in speech that are absent in writing (Castles et al., 2018). For example, gestures and facial expressions may be used along with a pronoun to indicate a referent in the animated movies, whereas a noun is necessary for the same referent in written texts.

Frequency and CD Distributions of the frequency (Zipf) and CD (logCD) are presented below in Fig. 1. The words’ Zipf and logCD are for lemmatized words. Frequency and CD measures for non-lemmatized words with POS tags can be found in the online database. The characters’ Zipf values ranged from 2.56 to 7.56 ($M = 4.27$, $SD = 1.02$). The words’ Zipfs ranged from 3.02 to 7.63 with the mean lying at 3.64 ($SD = .64$). The distributions, particularly for words, were heavily skewed to the left. The most frequent

Table 1 Distribution of syntactic categories of words in CCLOOW and CCLOWW G2

	Example	CCLOOW		CCLOWW G2	
		Type	Token	Type	Token
% Noun	书 <i>book</i>	53.34	18.36	49.05	22.65
% Verb	走 <i>walk</i>	28.09	25.67	31.11	24.14
% Adjective	红 <i>red</i>	8.57	5.60	10.12	5.88
% Adverb	还 <i>else</i>	3.89	13.46	4.78	12.48
% Interjection	啊 <i>ah</i>	2.46	2.41	.88	.50
% Measure word	个 (piece)	.78	2.28	.95	3.30
% Numeral	一百 <i>hundred</i>	.68	1.45	.62	2.90
% Onomatopoeia	哈哈 <i>haha</i>	.40	.07	.72	.16
% Pronoun	他 <i>he</i>	.35	12.56	.29	7.95
% Particle	所 (be)	.29	7.42	.22	2.99
% Preposition	从 <i>from</i>	.28	1.68	.39	2.93
% Conjunction	和 <i>and</i>	.22	1.27	.24	1.70
% Localizer	里 <i>in</i>	.16	.59	NA	NA
% Determiner	这 <i>this</i>	.15	1.69	.02	1.45
% Other		.34	5.49	.61	1.97

Note. The NA indicates that the figure was not reported/calculated in the original database

100 characters (2.55% of all types) and words (0.45% of all types) accounted for 58.40% and 56.63%, respectively, of all tokens. The pattern of CD distributions was similar. The average number of documents in which the words occurred was 11.32 ($SD = 26.09$). Over 50% of the words appeared in three or less documents. Two words occurred in all documents: the pronoun 你 (“you”) and the aspect marker 了. Similar to findings with previous lexical databases, the correlations between frequency and CD in CCLOOW are high (character $r = .95$; word $r = .92$).

Validating CCLOOW measures

Correlations with other Chinese lexical databases

We carried out Pearson’s correlational analyses on the frequency and CD measures of CCLOOW with that of two other Chinese lexical databases: CCLOWW G2 and SUBTLEX-CH. We selected the former because it was mostly based on picture books for children whose age is comparable to the target age of the current corpus. We selected the latter because it is the mostly widely used Chinese adult frequency norms and it was based on movie subtitles – a similar register to the language samples in CCLOOW. The analyses were based on 3,680 characters and 12,624 words shared across the three databases. The results are shown in Table 2.

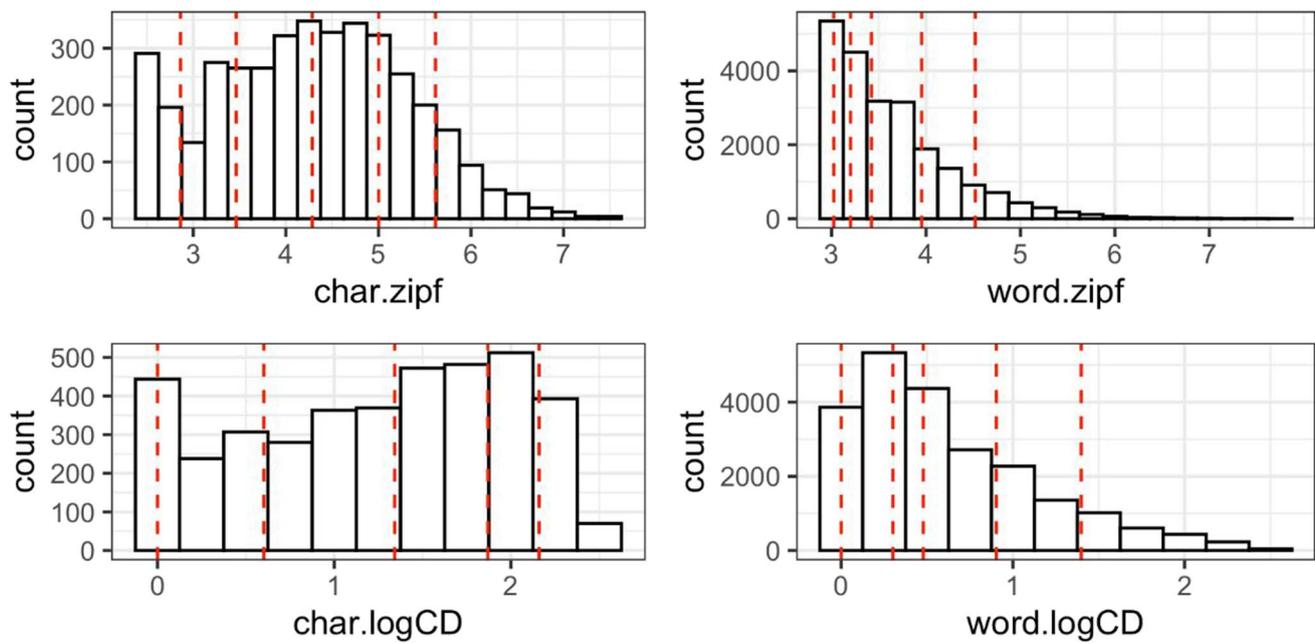


Fig. 1 Distributions of character and word frequency (Zipf) and CD (logCD) in CCLOOW. Dashed vertical lines indicate the 10, 25, 50, 75, and 90% percentiles

The correlations within databases were all very high ($r_s > .92$). Between-corpora correlations were also reasonably high, particularly for characters. Correlations between CCLOOW character measures and the other two databases were comparable (with CCLOWW G2: r range = .87–.88, with SUBTLEX-CH r range = .81–.88). In contrast, word measures in CCLOOW correlated better with that in CCLOWW G2 (r range = .74–.76) than with SUBTLEX-CH

(r range = .63–.71). This is consistent with Van Heuven et al.' (2014) finding that English children's written frequencies correlated with lexical frequencies based on children's TV channels than with adults' subtitle frequencies, highlighting the necessity for lexical databases made specific for children.

Table 2 Pearson's correlations between CCLOOW, CCLOWW G2, and SUBTLEX-CH

Measure	1	2	3	4	5
Character ($N = 3680$)					
1. Zipf_CCLOOW_	--				
2. Zipf_CCLOWW G2	.88	--			
3. LogFreq_SUBTLEX-CH	.86	.81	--		
4. LogCD_CCLOOW	.95	.86	.88	--	
5. LogCD_CCLOWW G2	.87	.97	.82	.88	--
6. LogCD_SUBTLEX-CH	.81	.77	.97	.88	.81
Word ($N = 12,624$)					
1. Zipf_CCLOOW	--				
2. Zipf_CCLOWW G2	.74	--			
3. LogFreq_SUBTLEX-CH	.68	.60	--		
4. LogCD_CCLOOW	.92	.73	.72	--	
5. LogCD_CCLOWW G2	.73	.95	.62	.76	--
6. LogCD_SUBTLEX-CH	.63	.56	.98	.71	.60

$p_s < .001$

Predicting children's character reading

Character naming We first validated the frequency and CD measures of CCLOOW characters using existing children's character naming data. We obtained the data of 52 Grades 2–3 children from Li et al. (2022) and collected additional data from 13 children of the same age. The children (age $M = 8.33$, $SD = .31$, 26 males) were tested individually in a quiet room at East China Normal University. They were instructed to read the characters presented on a paper sheet as accurately as they can. They were told that if they could not read a character, they could just skip it. The task took about 10 min. The participants all had normal or corrected-to-normal vision and were assessed by the teachers to have normal language and reading abilities. The experiments reported in this article have been approved by the East China Normal University Committee on Human Research Protection. Written consents were obtained from the children's guardians.

The analysis was based on 65 Grades 2–3 children's naming accuracy of 118 characters. Data were analyzed using

the packages *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) implemented in R (R Core Team, 2013). Two simple mixed effects models were built in which the characters' Zipf and logCD were separately added into the baseline model as a fixed factor. For this and the following regression analyses, following Barr et al.'s (2013) suggestion, we always started with a maximal random effects structure: $(1 + \text{Zipf}/\log\text{CD} \mid \text{subject}) + (1 + \text{Zipf}/\log\text{CD} \mid \text{item})$, and removed the random effect associated with the smallest variance one at a time to facilitate model convergence when a model could not converge. The final model structures are provided in the Appendix. Naming accuracies were binarily coded as the dependent variable in the logistic regression model. *P* values for the fixed effects were obtained using the *lmerTest* package. The effect of character Zipf was significant, $\beta = 4.77$, $SE = .46$, $z = 10.45$, $p < .001$, as well as the effect of character logCD, $\beta = 6.38$, $SE = .67$, $z = 9.46$, $p < .001$.

Predicting children's word reading

Next, to validate the words' frequency and CD measures in CCLOOW, we collected beginner readers' behavioral data in word naming and lexical decision experiments. Because we also want to compare its predictive validity with WF indexed from written corpora, we sampled words that did not correlate much on CCLOOW and CCLOWW G2 (Li et al., 2022) frequencies. The stimuli were 150 two-character words. The correlation between the words' frequencies in the two corpora were low, Pearson's $r = .21$, $p < .001$. The distributions of the target words' frequencies in the two databases are shown in Fig. 2.

Word naming Forty children participated in the experiment (18 males, age $M = 8.05$, $SD = .55$). One participant was excluded from the analysis due to technical error during data collection. Data from 39 children were included in the analysis. In the experiment, each child was seated in front of a computer in a quiet classroom on campus. They were presented a central fixation cross for 500 ms, followed by a target word at the center of a computer screen until an oral response was captured by the voice key or for a maximum of 3500 ms. The stimuli were presented in randomized order. The children were asked to say the word aloud to the microphone as quickly and accurately as they could. Optional breaks were provided after every 50 trials. The task took about 10 min.

Lexical decision Thirty-one children (13 males, age $M = 8.05$, $SD = .55$) participated in lexical decision. Two participants' data were excluded from the analysis because their accuracies were close to chance level (.50). Data from 29 children were included in the analysis. One hundred and fifty pseudowords were created by combing the first character of a target word with the second character of another one. Ten native adult Chinese speakers judged that they were made-up words. In the experiment, after a fixation cross of 500 ms, a target word or pseudoword was shown at the center of the screen for a maximum of 5000 ms. The children were asked to press one of two keys on the keyboard to indicate whether the word was a real word, meaning that they had seen it before and know its meaning, or a made-up word, meaning that they had not seen it before and did not know its meaning. The 300 stimuli were split into five lists of 30 target words and 30 pseudowords in each list. A blocked design was used. The assignment of the lists was counterbalanced across participants such that each list had similar total

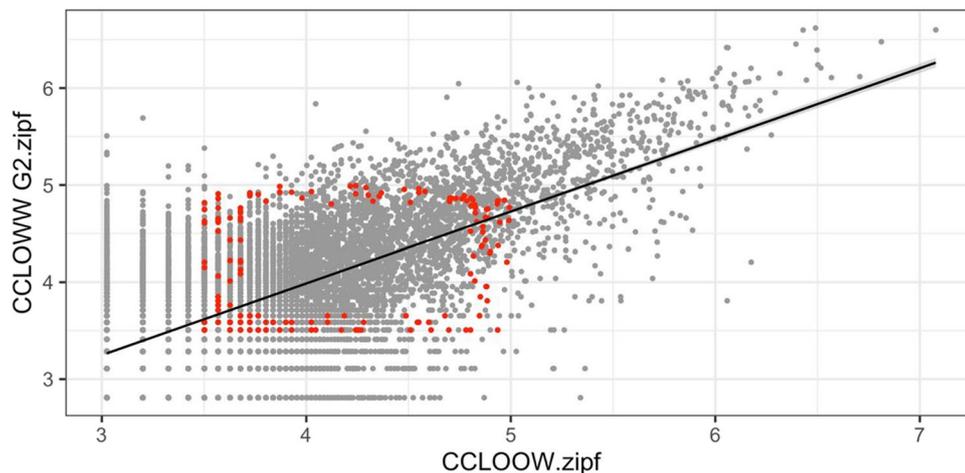


Fig. 2 Distributions of words' Zipf values in CCLOOW and in CCLOWW G2. *Grey dots* represent the common words in the two databases. *Colored dots* represent the sampled target words in the children's word naming and lexical decision experiments

number of occurrences in the five blocks. Optional breaks were provided between blocks. The task took about 20 min.

Results Data were cleaned before we proceeded to statistical analyses. In word naming, all data (5850 trials) were included in the analysis for naming accuracy; for naming RTs, incorrect responses, speech errors and responses with RTs beyond 3 standard deviations from the participant's mean were removed for each participant (16.74% of the data), resulting in 4870 valid observations. In lexical decision, 1.79% of the trials on the target words had RTs beyond 3 SDs of the participants' mean and were excluded, leading to 4272 valid observations in the analysis of accuracy; among them, 3533 trials were correctly responded to and the data were used in the analysis of lexical decision RTs. The average word naming accuracy was .90 ($SD = .30$) and the average naming RT was 966.09 ms ($SD = 295.85$ ms). The average lexical decision accuracy was .83 ($SD = .38$) and the average RT was 1126.31 ms ($SD = 491.91$ ms).

Again, we first built simple mixed effects models in which the words' Zipf and logCD computed in CCLOOW and CCLOWW G2 were separately added into the baseline model as a fixed factor. Because of high multicollinearity among the predictors, here we did not build full models with all variables included. Naming and lexical decision accuracy was binarily

coded as the dependent variable in logistic mixed effects models, and log-transformed RT was used as the dependent variable in linear mixed effects models. The results are shown in Table 3. Both WF and CD computed in the current database and in CCLOWW G2 significantly predicted the children's word naming and lexical decision accuracies and RTs.

Next, to examine the unique effects of each frequency measure, we constructed multiple logistic and linear mixed effects models in a stepwise manner. The base model first included some psycholinguistic variables known to influence naming and lexical decision performances, including orthographic complexity (i.e., number of strokes), concreteness, age of acquisition (AoA), the number of strokes and the phonetic regularity of the first and second constituting characters (C1 and C2). Phonetic regularity was coded to indicate non-phonograms (0) and whether (1) or not (-1) a phonogram is identical to its phonetic radical in pronunciation. Concreteness values were obtained for all 150 target words from Xu and Li (2020). The numbers of strokes for the 150 words and the phonetic regularity for 104 of the first and for 102 of the second constituting characters were acquired from Liu et al. (2007). Nonetheless, among the variables, only AoA significantly predicted the children's naming and lexical decision. Therefore, we only kept AoA as the fixed effect in the base model (Model 0).

Table 3 Results of simple mixed-effects logistic and linear models fitted to accuracy and log RT in naming and lexical decision

	Accuracy				RT			
	β	<i>SE</i>	<i>z</i>	<i>p</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>
Word naming								
Zipf_CCLOOW	1.09	.33	3.27	.001	-.08	.02	-3.38	< .001
M/C R^2	.042/.594				.020/.563			
logCD_CCLOOW	1.65	.41	4.01	< .001	-.12	.03	-4.31	< .001
M/C R^2	.061/.591				.048/.559			
Zipf_CCLOWW G2	.77	.32	2.39	.017	-.05	.02	-2.55	.012
M/C R^2	.023/.592				.011/.565			
logCD_CCLOWW G2	.96	.35	2.72	.007	-.08	.02	-3.37	< .001
M/C R^2	.030/.591				.020/.568			
Lexical decision								
Zipf_CCLOOW	1.13	.23	4.87	< .001	-.07	.03	-2.57	.014
M/C R^2	.096/.468				.007/.366			
logCD_CCLOOW	1.59	.29	5.54	< .001	-.09	.03	-2.77	.008
M/C R^2	.117/.467				.008/.363			
Zipf_CCLOWW G2	1.22	.25	4.87	< .001	-.06	.02	-3.47	< .001
M/C R^2	.070/.504				.005/.343			
logCD_CCLOWW G2	1.28	.30	4.25	< .001	-.06	.02	-3.05	.003
M/C R^2	.063/.512				.005/.346			

Note. M/C R^2 = Marginal/Conditional R^2

In the next steps, CCLOWW G2 word frequency was entered (Model 1), followed by CCLOWW G2 frequency of the constituting characters (C1_Zipf_CCLOWW G2 and C2_Zipf_CCLOWW G2, Model 2). Then we entered CCLOWW word frequency (Model 3) and finally CCLOWW character frequency (Model 4). Changes in the Akaike information criterion (AIC) were used to identify the set of predictors that maximize model fit. Likelihood ratio test was used to compare the nested models, the results of which are summarized in Table 4. Model 3 fit the data better than other models across all outcome variables.

Results of the final regression models, this is, Model 3, are presented in Table 5. CCLOWW WF was a significant predictor of the children’s word recognition in all outcome measures. Whereas the effect of CCLOWW G2 written word frequency became not significant once CCLOWW WF was included in the model. CCLOWW G2 written frequency of the constituting characters, particularly the first character, also consistently predicted word recognition performance.

We then examined the additional effect of subtitle-based adult frequencies on the children’s data by adding SUBTLEX-CH frequencies to the above multiple regression models. It was a significant predictor of children’s naming ($\beta = -.02, SE = .01, t = -2.77, p = .006$) and lexical decision RTs ($\beta = .03, SE = .01, t = 2.57, p = .001$) but not significant on naming ($\beta = -.04, SE = .28, z = .13, p = .894$) or lexical decision accuracies ($\beta = .10, SE = .22, z = .347, p = .638$). Intriguingly, its effect on lexical decision RTs was in the reverse direction of the typically WF effect. Therefore, we built another simple mixed-effects model keeping only SUBTLEX-CH frequency as the fixed effect in analyzing the data. The effect of SUBTLEX-CH frequencies was only approaching significant on lexical decision RTs ($\beta = -.03, SE = .01, t = -1.90, p = .060$).

Predicting adults’ word naming and lexical decision RTs

Given Li et al.’s (2022) finding that frequency measures based on children’s books explained variance in adults’ word naming and lexical decision RTs in addition to adult frequency measures, we also explored whether CCLOWW frequencies can explain adults’ word recognition. We obtained adult word naming data from Li et al. (2022) and lexical decision data from Tsang et al. (2018) for 277 two-character words and regressed the RTs on SUBTLEX-CH frequencies and CCLOWW frequencies. Accuracies were not analyzed because they were at ceiling. A number of other significant predictors of the RTs found in Li et al. (2022) were also included in the linear regression models: frequency of the first character in CCLOWW (C1_Zipf_CCLOWW), frequency of the second character in CCLOWW (C2_Zipf_CCLOWW), total number of strokes (*N* strokes), regularity of the first character (Regularity_C1), number of words the first (Ortho *N*_C1) and the second character occurs in (Ortho *N*_C2), number of characters the semantic radical of the first (SR ortho *N*_C1) and the second character occurs in (SR ortho *N*_C2), number of characters the phonetic radical of the first (PR ortho *N*_C1) and the second character occurs in (PR ortho *N*_C2). These variables were obtained from Sun et al. (2018). Concreteness obtained from Xu and Li (2020) was also included in the lexical decision model. AoA and the stroke counts of constituting characters were not included to avoid multicollinearity issues. The results are shown in Table 6.

The significant predictors jointly explained 60.48% of the variance in adults’ word naming RTs and 57.42% of the variance in lexical decision RTs. The effect of CCLOWW word frequencies was significant on lexical decision but not on

Table 4 Results of likelihood ratio test for model comparisons

	Naming accuracy	Naming RT	Lexical decision accuracy	Lexical decision RT
Model 0: AoA	AIC = 2189.6	AIC = - 1389.0	AIC = 2524.8	AIC = 2235.2
Model 1: AoA + Zipf_CCLOWW G2	$\chi^2 = 3.09, df = 1, p = .079, AIC = 2188.5$	$\chi^2 = 2.20, df = 1, p = .138, AIC = -1351.5$	$\chi^2 = 5.94, df = 1, p = .015, AIC = 2520.9$	$\chi^2 = 6.44, df = 1, p = .011, AIC = 2230.8$
Model 2: AoA + Zipf_CCLOWW G2 + C1_Zipf_CCLOWW G2 + C2_Zipf_CCLOWW	$\chi^2 = 40.38, df = 2, p < .001, AIC = 2152.1$	$\chi^2 = 22.69, df = 2, p < .001, AIC = -1407.9$	$\chi^2 = 9.20, df = 2, p = .010, AIC = 2525.7$	$\chi^2 = 13.60, df = 2, p = .001, AIC = 2221.2$
Model 3: AoA + Zipf_CCLOWW G2 + C1_Zipf_CCLOWW G2+ C2_Zipf_CCLOWW + C1_Zipf_CCLOWW G2+ C2_Zipf_CCLOWW	$\chi^2 = 5.95, df = 1, p = .015, AIC = 2148.2$	$\chi^2 = 7.49, df = 1, p = .006, AIC = -1413.3$	$\chi^2 = 12.87, df = 1, p < .001, AIC = 2504.8$	$\chi^2 = 11.04, df = 1, p < .001, AIC = 2212.1$
Model 4: AoA + Zipf_CCLOWW G2 + C1_Zipf_CCLOWW G2+ C2_Zipf_CCLOWW + C1_Zipf_CCLOWW G2+ C2_Zipf_CCLOWW	$\chi^2 = 1.33, df = 2, p = .515, AIC = 2150.8$	$\chi^2 = .90, df = 2, p = .635, AIC = -1410.2$	$\chi^2 = 3.86, df = 2, p = .145, AIC = 2505.0$	$\chi^2 = 3.16, df = 2, p = .206, AIC = 2212.9$

Note. C1 first character, C2 second character

Table 5 Results of final multiple regression models fitted to accuracy and log RT in naming and lexical decision

	Naming accuracy					Naming RT				
	β	<i>SE</i>	<i>z</i>	<i>p</i>	VIF	β	<i>SE</i>	<i>t</i>	<i>p</i>	VIF
(Intercept)			-2.94	.003				36.92	< .001	
AoA	-.20	.09	-2.30	.022	1.33	.02	.01	2.18	.032	1.29
Zipf_CLOWW G2	-.15	.32	-.46	.646	1.46	.001	.03	.05	.962	1.39
Zipf_CLOWW	.70	.28	2.47	.013	1.12	-.06	.02	-2.72	.008	1.11
C1_Zipf_CLOWW G2	1.23	.20	6.20	< .001	1.07	-.08	.02	-4.60	< .001	1.04
C2_Zipf_CLOWW G2	.61	.21	2.93	.003	1.17	-.03	.02	-1.67	.097	1.16
Marginal R^2 / Conditional R^2	.207/ .550					.057/ .558				
	Lexical decision accuracy					Lexical decision RT				
(Intercept)			-1.68	.093				47.10	< .001	
AoA	-.28	.07	-3.99	< .001	1.33	.01	.01	2.28	.025	1.24
Zipf_CLOWW G2	.25	.24	1.04	.300	1.41	-.02	.02	-1.06	.294	1.37
Zipf_CLOWW	.83	.22	3.74	< .001	1.12	-.06	.02	-3.32	.001	1.10
C1_Zipf_CLOWW G2	.41	.15	2.64	.008	1.07	-.05	.01	-3.48	.001	1.03
C2_Zipf_CLOWW G2	.24	.16	1.43	.153	1.19	-.02	.01	-1.44	0.152	1.14
Marginal R^2 / Conditional R^2	.144/ .463					.019/ .350				

Significant *p* values smaller than .05 are highlighted in bold. C1 first character, C2 second character

naming RTs. We then explored the unique variance in the RTs explained by SUBTLEX-CH frequencies, CCLOWW frequencies of the words and of the constituting characters in stepwise multiple regression analyses. A base model with the significant predictors in the previous analyses other than the frequency measures were first built (Step 1). We entered SUBTLEX-CH word frequencies in Step 2 and CCLOWW word frequencies in Step 3. The CCLOWW character frequencies of the first and the second constituting characters were entered in Steps 4 and 5 respectively. The results (Table 7) show that in word naming, the effect of CCLOWW words' Zipf was significant, $\beta = -.06$, $SE = .03$, $t = -2.02$, $p = .043$, and contributed 15.40% variance in naming RTs.

Nevertheless, its effect became nonsignificant once the first characters' CCLOWW Zipf were added, which contributed 22.00% additional variance. In comparison, in lexical decision, the effect of CCLOWW words' Zipf was consistently significant and explained 8.12% variance while characters' Zipfs jointly explained 6.70% variance in the RTs.

Discussion and conclusion

CCLOWW is the first lexical database based on animated movies and TV series for Chinese children. Compared with the recent CLOWW database (Li et al., 2022) compiled from printed books, CCLOWW provides frequency and contextual diversity values of characters and words based on another important source of language input in children's environment. The database is particularly appropriate for profiling the language experience of children transitioning from

pre-readers to beginner readers (3–9-year-olds). CCLOWW will add to a comprehensive description of the lexical statistics for developing language learners in mainland China.

Our analysis shows that frequencies computed from animated movies and TV series for children are reliable estimates of young children's experience with characters and words. The frequency and contextual diversity measures correlated well within the CCLOWW database. Their correlations with the children's lexical database based on printed books (CLOWW G2, Li et al., 2022) and with the adult subtitle database (SUBTLEX-CH, Cai & Brysbaert, 2010) were also reasonably high. In particular, the correlations of word measures were higher between the two children's corpora than between the children's and the adult corpora. This indicates that language materials for children, regardless of the register, differ from that for adults, highlighting the importance of building specialized corpora for children. This suggestion was corroborated by the results of the word recognition experiments with Grade 2 children.

CCLOWW measures significantly predicted Grade 2 children's character naming, word naming and lexical decision performance. Particularly, the WF effects on word recognition were robust and beyond other frequency measures (CLOWW G2 and SUBTLEX-CH). When WF measures from these databases were submitted to analysis in one model, only CCLOWW WF remained significant. This contrast shows that beginner readers' lexical processing is much better explained by frequencies computed from children's materials, particularly spoken texts, than by that from adults' language samples, again stressing the necessity of

Table 6 Results of linear regression models predicting adults' word naming and lexical decision RTs

Predictor	β	SE	t	p	VIF
Word naming RT					
Zipf_CCLOOW	.01	.02	.59	.559	1.65
LogFreq_SUBTLEX-CH	-.21	.02	-12.85	< .001	1.42
N strokes	.02	.01	9.23	< .001	1.21
C1_ Zipf_CCLOOW	-.46	.02	-24.72	< .001	2.17
C2_ Zipf_CCLOOW	.04	.02	2.30	.022	2.16
Regularity_C1	.04	.02	2.30	.022	1.22
Ortho N_C1	.01	.01	8.56	< .001	2.34
Ortho N_C2	.01	.01	2.43	.015	1.88
SR ortho N_C1	-.01	.01	-14.40	< .001	1.17
SR ortho N_C2	.01	.01	12.57	< .001	1.19
PR ortho N_C1	.01	.01	5.91	< .001	1.45
PR ortho N_C2	-.02	.01	-10.29	< .001	1.42
Adjusted $R^2 = .605$					
Lexical decision RT					
Zipf_CCLOOW	-.13	.016	-8.497	< .001	1.98
LogFreq_SUBTLEX-CH	-.21	.01	-19.79	< .001	1.90
N strokes	.01	.01	11.33	< .001	1.24
C1_ Zipf_CCLOOW	-.13	.01	-12.08	< .001	2.19
C2_ Zipf_CCLOOW	.07	.01	6.40	< .001	2.20
Regularity_C1	.04	.01	6.72	< .001	1.23
Ortho N_C1	.01	.01	5.07	< .001	2.39
Ortho N_C2	-.01	.01	-6.19	< .001	1.88
SR ortho N_C1	-.01	.01	-3.14	.002	1.18
SR ortho N_C2	.01	.01	3.97	< .001	1.22
PR ortho N_C1	.01	.01	3.15	.002	1.47
PR ortho N_C2	-.01	.01	-11.40	< .001	1.46
Concreteness	.04	.01	4.64	< .001	1.63
Adjusted $R^2 = .574$					

Note. C1_ Zipf_CCLOOW = CCLOOW frequency of the first character; C2_ Zipf_CCLOOW = CCLOOW frequency of the second character; N strokes = stroke counts of word; Regularity_C1 = regularity of the first character ;Ortho N_C1/ Ortho N_C2 = number of words the character occurs in; SR ortho N_C1/SR ortho N_C2 = number of characters the semantic radical of the character occurs in; PR ortho N_C1/PR ortho N_C2 = number of characters the phonetic radical of the character occurs in

children's lexical databases in developmental psycholinguistic research.

The finding that the effect of CCLOWW G2 written WF was also diminished once CCLOOW WF was added in the models implicated the possibility that beginner reader's lexical knowledge is still largely affected by their spoken language environment. This suggestion was also supported by our finding that CCLOWW G2 frequency of the words' constituting characters did predict the children's word recognition performance. For Chinese children, the priority of the early school years is to learn to read characters and not words. Shu et al. (2003) has similarly suggested that the task of learning characters is the heaviest for children until Grade 3. Therefore, the beginner readers in our word naming and lexical decision experiments might have been intensively exposed to written characters rather than written words, thus demonstrating a strong frequency effect of written characters and not words. It is likely that the written WF effect would be more pronounced in older children and children who started reading early. Whereas CCLOOW WF might perform even better with younger children with limited reading experience. We aim to investigate these questions in future work. Note that children in rural areas might not have as much access to TV and screen media as children living in cities. Given that they are also likely exposed to written texts later or less, other sources of language input such as daily conversation might have a greater impact. How CCLOOW statistics might affect character and word reading of these children also awaits investigation.

In our analysis with the adults' word naming and lexical decision data, the effect of CCLOOW word frequencies was diminished once character frequencies were added in the model predicting naming RTs, whereas the WF effect remained robust on lexical decision. Moreover, CCLOOW character frequencies explained as much as 22.20% extra variance in naming RTs, in addition to adults' word frequencies and CCLOOW word frequencies. Whereas in lexical decision RTs, CCLOOW word frequencies contributed more variance (8.12%) than character frequencies (6.70%). Intriguingly, CCLOOW word frequencies contributed three

Table 7 Results of stepwise regression analyses of adults' word naming and lexical decision RTs

Step	Predictor	Word naming RT		Lexical decision RT	
		Adjusted R^2	ΔR^2	Adjusted R^2	ΔR^2
1	Lexical & sublexical variables	.180		.133	
2	LogFreq_SUBTLEX-CH	.229	.049***	.426	.293***
3	Zipf_CCLOOW	.383	.154*	.507	.081***
4	C1_ Zipf_CCLOOW	.603	.220***	.559	.052***
5	C2_ Zipf_CCLOOW	.605	.002*	.574	.015***

* $p < .05$. *** $p < .001$

times more variance (15.40%) than SUBTLEX-CH frequencies (4.93%) in naming, but the pattern was reversed for lexical decision. These results were consistent with Li et al.'s (2022) finding, and they together suggest that early language experience in childhood might have lasting impacts on the mature lexicon, particularly early spoken language experience. How early spoken and written language input variations might impact lexical processing in skilled language users warrant further investigations in the future.

Although frequencies are the most established lexical predictors of lexical processing, it should be noted that there are other indices of words' occurrences in spoken and written texts which likely influence lexical representation and acquisition. We have shown that CCLOOW CD could also predict children's word and character reading, but the effects were not compared to that of frequency. The two indices were often separately analyzed due to high correlation, yet it is possible to disentangle their effects by examining an alternative measure of CD – semantic diversity (SD). SD considers the content rather than the number of documents in which word occurs to capture context-dependent variability in word meaning (Hoffman et al., 2013; Jones et al., 2012). Importantly, some has found that SD outperformed WF and CD in predicting naming and lexical decision latencies in adults (Chang & Lee, 2018; Hoffman et al., 2013; Jones et al., 2012; Jones et al., 2017) and children (Hsiao & Nation, 2018). Future work should also consider indices such as SD to measure how variability in words' meaningful occurrences in context influences children's reading and learning to read in Chinese.

The CCLOOW database is accessible online at <https://www.learn2read.cn/ccloow>. Statistics of frequency, contextual diversity, word length, syntactic category are provided for characters, non-lemmatized POS-tagged words, and lemmatized words. Search functions are made available to tailor to the needs of users. In the future, we plan to include other lexical and sublexical variables to provide a comprehensive account of the characters and words in the language environment of young Chinese children from early to late childhood.

Appendix. Statistical model structures

1. Children's character naming

Maximal model/final model:

$$\text{ACC} \sim \text{Zipf}/\log\text{CD} + (1 + \text{Zipf}/\log\text{CD} \mid \text{subject}) + (1 + \text{Zipf}/\log\text{CD} \mid \text{item})$$

2. Children's word naming/lexical decision

2.1 Simple regressions (Table 3)

Maximal model:

$$\text{ACC}/\log\text{RT} \sim \text{Zipf}/\log\text{CD} + (1 + \text{Zipf}/\log\text{CD} \mid \text{subject}) + (1 + \text{Zipf}/\log\text{CD} \mid \text{item})$$

2.2 Stepwise multiple regressions (Table 5)

Base model (Step 1):

$$\text{ACC}/\log\text{RT} \sim \text{AoA} + \text{N.Strokes} + \text{concreteness} + \text{C1_regularity} + \text{C2_regularity} + \text{C1.stroke} + \text{C2.stroke} + \text{C1FamilySize} + \text{C2FamilySize} + \text{C1SRFamilySize} + \text{C2SRFamilySize} + \text{C1PRFamilySize} + \text{C2PRFamilySize} + (1 \mid \text{subject}) + (1 \mid \text{item})$$

Final model:

$$\text{ACC}/\log\text{RT} \sim \text{AoA} + \text{Zipf_CCLOWW G2} + \text{Zipf_CCLOWW} + \text{C1_Zipf_CCLOWW G2} + \text{C2_Zipf_CCLOWW G2} + (1 \mid \text{subject}) + (1 \mid \text{item})$$

3. Adults' word naming/lexical decision (Table 6)

Final model:

$$\text{ACC}/\log\text{RT} \sim \text{Zipf_CCLOWW} + \text{LogFreq_SUBTLEX-CH} + \text{C1_Zipf_CCLOWW} + \text{C2_Zipf_CCLOWW} + \text{N.Strokes} + \text{concreteness} + \text{C1_regularity} + \text{C2_regularity} + \text{C1FamilySize} + \text{C2FamilySize} + \text{C1SRFamilySize} + \text{C2SRFamilySize} + \text{C1PRFamilySize} + \text{C2PRFamilySize} + (1 \mid \text{subject}) + (1 \mid \text{item})$$

Note. C1FamilySize = Ortho N_{C1} , C2FamilySize = Ortho N_{C2} , C1SRFamilySize = SR Ortho N_{C1} ; C2SRFamilySize = SR Ortho N_{C2} , C1PRFamilySize = PR Ortho N_{C1} , C2PRFamilySize = PR Ortho N_{C2}

Funding This work was supported by the National Natural Science Foundation of China (grant numbers 31970987 to QC; 32100857 to JW), the Basic Research Project of Shanghai Science and Technology Commission (grant No.19JC1410101), the Open Research Fund of Shanghai Key Laboratory of Brain Functional Genomics (East China Normal University) and the Fundamental Research Funds for the Central Universities (grant No. 2022ECNU-YYJ021 to LL).

Declarations

Conflict of interest The authors declare no conflicts of interest.

Ethics approval The study has been approved by the East China Normal University Committee on Human Research Protection. Written consents have been obtained from the child participants' parents/guardians.

References

- Adelman, J., Brown, G., & Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 19(9), 814–823. <https://doi.org/10.1111/j.1467-9280.2006.01787.x>
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable

- words. *Journal of Experimental Psychology: General*, 133(2), 283–316. <https://doi.org/10.1037/0096-3445.133.2.283>
- Balota, David A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3), 445–459. <https://doi.org/10.3758/BF03193014>
- Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Brysaert, M., & Ellis, A. W. (2016). Aphasia and age-of-acquisition: Are early-learned words more resilient? *Aphasiology*, 30(11), 1240–1263. <https://doi.org/10.1080/02687038.2015.1106439>
- Brysaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. <https://doi.org/10.3758/BRM.41.4.977>
- Brysaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58(5), 412–424. <https://doi.org/10.1027/1618-3169/a000123>
- Brysaert, M., Mandler, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27(1), 45–50. <https://doi.org/10.1177/0963721417727521>
- Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, and Computers*, 30(2), 272–277. <https://doi.org/10.3758/BF03200655>
- Cai, Q., & Brysaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, 5(6), e10729. <https://doi.org/10.1371/journal.pone.0010729>
- Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction-based analysis of child directed speech. *Cognitive Science*, 27, 843–873. <https://doi.org/10.1016/j.cogsci.2003.06.001>
- Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, 19(1), 5–51. <https://doi.org/10.1177/1529100618772271>
- Chang, Y.-N., & Lee, C.-Y. (2018). Semantic ambiguity effects on traditional Chinese character naming: A corpus-based approach. *Behavior Research Methods*, 50(6), 2292–2304. <https://doi.org/10.3758/s13428-017-0993-4>
- Chen, Q., Huang, X., Bai, L., Xu, X., Yang, Y., & Tanenhaus, M. K. (2017a). The effect of contextual diversity on eye movements in Chinese sentence reading. *Psychonomic Bulletin and Review*, 24(2), 510–518. <https://doi.org/10.3758/s13423-016-1119-1>
- Chen, Q., Zhao, G., Huang, X., Yang, Y., & Tanenhaus, M. K. (2017b). The effect of character contextual diversity on eye movements in Chinese sentence reading. *Psychonomic Bulletin and Review*, 24(6), 1971–1979. <https://doi.org/10.3758/s13423-017-1278-8>
- Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32, 133–143.
- Dore, R. A., Logan, J., Lin, T. J., Purtell, K. M., & Justice, L. (2020). Characteristics of children's media use and gains in language and literacy skills. *Frontiers in Psychology*, 11, 2224. <https://doi.org/10.3389/FPSYG.2020.02224/BIBTEX>
- Geng, Z., Yan, H., Qiu, X., & Huang, X. (2020). fastHan: A BERT-based multi-task toolkit for Chinese NLP. *ArXiv Preprint*, arXiv:2009.08633.
- Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, 48(3), 963–972. <https://doi.org/10.3758/S13428-015-0621-0>
- Gong, X., Lu, Y., Zhou, Z., & Qian, Y. (2022). Layer-wise fast adaptation for end-to-end multi-accent speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2842–2853. <https://doi.org/10.21437/Interspeech.2021-1075>
- Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*, 41(4), 976–995. <https://doi.org/10.1111/COGS.12392>
- Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, 45, 718–730. <https://doi.org/10.3758/s13428-012-0278-x>
- Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language*, 103, 114–126. <https://doi.org/10.1016/j.jml.2018.08.005>
- Huang, X., Lin, D., Yang, Y., Xu, Y., Chen, Q., & Tanenhaus, M. K. (2020). Effects of character and word contextual diversity in Chinese beginning readers. *Scientific Studies of Reading*. <https://doi.org/10.1080/10888438.2020.1768258>
- Johns, B. T., & Jones, M. N. (2022). Content matters: Measures of contextual diversity must consider semantic content. *Journal of Memory and Language*, 123, 104313. <https://doi.org/10.1016/J.JML.2021.104313>
- Jones, M., Johns, B., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, 66(2), 115–124. <https://doi.org/10.1037/a0026727>
- Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an Organizing Principle of the Lexicon. *Progress in Brain Research*, 232, 239–283. <https://doi.org/10.1016/bs.plm.2017.03.008>
- Joseph, H., & Nation, K. (2017). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology*, 166, 190–211. <https://doi.org/10.1016/j.jecp.2017.08.010>
- Keuleers, E., Brysaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3), 643–650. <https://doi.org/10.3758/BRM.42.3.643>
- Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. RI: Brown University Press.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lampe, L., Hameau, S., & Nickels, L. (2021). Semantic variables both help and hinder word production: Behavioral evidence from picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(1), 72–97. <https://doi.org/10.1037/xlm0001050>
- Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, and Computers*, 36(1), 156–166. <https://doi.org/10.3758/BF03195560>
- Li, P., & Jeong, H. (2020, December 1). The social brain of language: grounding second language learning in social interaction. *NPI Science of Learning*. Springer Nature. <https://doi.org/10.1038/s41539-020-0068-7>
- Li, L., Yang, Y., Song, M., Zhang, M. Y., Chen, Q. R., & Cai, Q. (2022). CLOWW: A grade-level Chinese children's lexicon of written words. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-01890-9>

- Linebarger, D. L., & Vaala, S. E. (2010). Screen media and language development in infants and toddlers: An ecological perspective. *Developmental Review, 30*(2), 176–202. <https://doi.org/10.1016/j.DR.2010.03.006>
- Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39*(2), 192–198. <https://doi.org/10.3758/BF03193147>
- Liu, Q., Chen, Z., Li, H., Huang, M., & Lu, Y. (2020). Modular end-to-end automatic speech recognition framework for acoustic-to-word model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28*, 2174–2183.
- Madigan, S., Browne, D., Racine, N., & Mori, C. (2019). Association between screen time and children's performance on a developmental screening test. *JAMA Pediatrics, 173*(3), 244–250. <https://doi.org/10.1001/jamapediatrics.2018.5056>
- Masterson, J., Stuart, M., Dixon, M., Lovejoy, D., & Lovejoy, S. (2003). *The Children's Printed Word Database*. Available online at: www.essex.ac.uk/psychology/cpwd
- Masterson, J., Stuart, M., & Dixon, M. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology, 101*(2), 221–242. <https://doi.org/10.1348/000712608X371744>
- Mayer, R. E., Moreno, R., Boire, M., & Vagge, S. (1999). Maximizing Constructivist Learning from Multimedia Communications by Minimizing Cognitive Load. *Journal of Educational Psychology, 91*(4), 638–643. <https://doi.org/10.1037/0022-0663.91.4.638>
- Ministry of Education, R. O. C. (2013). *Table of General Standard Chinese Characters*. Available online at: http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm
- 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>
- Myers, L. J., LeWitt, R. B., Gallo, R. E., & Maselli, N. M. (2017). Baby FaceTime: can toddlers learn from online video chat? *Developmental Science, 20*(4), e12430. <https://doi.org/10.1111/DESC.12430>
- Pagan, A., & Nation, K. (2019). Learning words via reading: Contextual diversity, spacing, and retrieval effects in adults. *Cognitive Science, 43*, 1–24. <https://doi.org/10.1111/cogs.12705>
- Pagani, L., Fitzpatrick, C., & Barnett, T. (2013). Early childhood television viewing and kindergarten entry readiness. *Pediatric Research, 74*(3), 350–355. <https://doi.org/10.1038/pr.2013.105>
- Perea, M., Soares, A. P., & Comesaña, M. (2013). Contextual diversity is a main determinant of word identification times in young readers. *Journal of Experimental Child Psychology, 116*, 37–44. <https://doi.org/10.1016/j.jecp.2012.10.014>
- Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 275–283. <https://doi.org/10.1037/a0034058>
- R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Rosa, E., Tapia, J. L., Perea, M., Smith, L., Simões, A., & Almeida, J. de. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS ONE, 12*(6), e0179004. <https://doi.org/10.1371/journal.pone.0179004>
- Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83*(5), 1762–1774. <https://doi.org/10.1111/j.1467-8624.2012.01805.x>
- Schroeder, S., Würzner, K. M., Heister, J., Geyken, A., & Kliegl, R. (2014). childLex: a lexical database of German read by children. *Behavior Research Methods, 47*(4), 1085–1094. <https://doi.org/10.3758/s13428-014-0528-1>
- Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of school Chinese: Implications for learning to read. *Child development, 74*(1), 27–47. <https://doi.org/10.1111/1467-8624.00519>
- Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, Á., ... & Comesaña, M. (2014). ESCOLEX: A grade-level lexical database from European Portuguese elementary to middle school textbooks. *Behavior Research Methods, 46*, 240–253. <https://doi.org/10.3758/s13428-013-0350-1>
- Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (CLD): A large-scale lexical database for simplified Mandarin Chinese. *Behavior Research Methods, 50*(6), 2606–2629. <https://doi.org/10.3758/s13428-018-1038-3>
- Tan, L., & Perfetti, C. A. (1999). Phonological activation in visual identification of Chinese two-character words for a review of recent literature. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(2), 382–393. <https://doi.org/10.1037/0278-7393.25.2.382>
- Taylor, K., Devereux, B., Acres, K., Randall, B., & Tyler, L. (2012). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition, 122*, 363–374. <https://doi.org/10.1016/j.cognition.2011.11.001>
- Terzopoulos, A. R., Duncan, L. G., Wilson, M. A. J., Niolaki, G. Z., & Masterson, J. (2017). HelexKids: A word frequency database for Greek and Cypriot primary school children. *Behavior Research Methods, 49*(1), 83–96. <https://doi.org/10.3758/s13428-015-0698-5>
- Tsang, Y. K., Huang, J., Lui, M., Xue, M., Chan, Y. W. F., Wang, S., & Chen, H. C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods, 50*(5), 1763–1777. <https://doi.org/10.3758/s13428-017-0944-0>
- Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. <https://doi.org/10.1080/17470218.2013.850521>
- Xu, X., & Li, J. (2020). Concreteness/abstractness ratings for two-character Chinese words in MELD-SCH. *PLoS ONE, 15*(6), e0232133. <https://doi.org/10.1371/journal.pone.0232133>
- Xue, N., Xia, F., Chiou, F., & Palmer, M. (2019). The Penn Chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering, 11*(2), 207–238. <https://doi.org/10.1017/S135132490400364X>
- Yap, M., Tse, C., & Balota, D. A. (2009). Individual differences in the joint effects of semantic priming and word frequency revealed by RT distributional analyses: The role of lexical integrity. *Journal of Memory and Language, 61*, 303–325. <https://doi.org/10.1016/j.jml.2009.07.001>

Open practices statement The database is available at <https://www.learn2read.cn/cclow>. Experimental data collected in this study and analysis script are provided at <https://osf.io/pyc6k/>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.