Behavior Research Methods (2023) 55:1874–1889
https://doi.org/10.3758/s13428-022-01890-9

CCLOWW: A grade-level Chinese children's lexicon of written words

Luan Li[1] · Yang Yang[1] · Ming Song[1] · Siyi Fang[1] · Manyan Zhang[1] · Qingrong Chen[2] · Qing Cai[1,3]

Accepted: 24 May 2022 / Published online: 1 July 2022

Abstract

In this article, we present the Chinese Children's Lexicon of Written Words (CCLOWW), the first grade-level database that provides frequency statistics of simplified Chinese characters and words for children. The database computes from a corpus of 34,671,424 character tokens and 22,427,010 word tokens (including single- and multicharacter words), extracted from 2131 books. It contains 6746 different character types and 153,079 different word types. CCLOWW provides several frequency indices of simplified Chinese for three grade levels (grade 2 and below, grades 3–4, grades 5–6) to profile children's experience with written Chinese in and outside of school. We describe in this article the distributions of frequency and contextual diversity of the characters and words, as well as word length and syntactic categories of the words in the corpus and the subcorpora. We also report results of correlation analyses with other written corpora and of several naming and lexicon decision experiments. The findings suggest that CCLOWW frequency measures correlate well with other corpora. Importantly, they could reliably predict children's and adults' naming and lexical decision performances. They could also explain variance in adults' visual word recognition, in addition to frequency measures computed in an adult corpus, indicating that early print exposure might influence readers' lexical processing later on beyond an age of acquisition effect. CCLOWW will help researchers in language processing and development as well as educators with selecting language materials appropriate for children's developmental stages. The database is freely available online at https://www.learn2read.cn/database/.

Keywords Lexical database · Children · Chinese · Frequency · Contextual diversity

Introduction

Lexical databases are important tools for selecting experimental materials for research on language and reading processes. This is because the occurrence of words in corpora is suggested to reflect the population's average experiences with words and thus affect the processing of them. For example, one of the most established predictors of visual word recognition—word frequency—is an index based on objective word counts in corpora that sampled large numbers of reading materials. Research has consistently shown a robust frequency effect in written word processing; that is, high-frequency words enjoy a processing advantage such that they are recognized more quickly and accurately than low-frequency words (for a review, see Marc Brysbaert et al., 2018). Estimates of word occurrence in databases are thus regarded as reliable reflections of the population's average experience with the words. As such, studies probing into the mechanisms of mental representation and processing of written words have been relying on frequency indices from databases to construct experimental stimuli.

Many written word databases have been made available for a variety of languages (e.g., Keuleers et al., 2010; Soares et al., 2014; Tse & Yap, 2017; Van Heuven et al., 2014). Nevertheless, they have been developed mainly to contribute to research with skilled readers. These corpora cannot adequately account for children's reading experiences because they reflect an established and relatively stable system and do not capture the dynamics of a developing one. Children's vocabulary and reading abilities are not as adequate as adults'; how they read and learn to read are also undergoing

✉ Qing Cai
qcai@psy.ecnu.edu.cn

[1] Key Laboratory of Brain Functional Genomics (MOE & STCSM), Institute of Brain and Education Innovation, School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

[2] School of Psychology, Nanjing Normal University, Nanjing, China

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

dramatic changes. At different developmental stages, some processes underlying reading and reading acquisition might be more salient than others. For example, according to Ehri's (2005) phase theory, in the very early stages of reading development, children use their partial knowledge of the grapheme-phoneme relations to associate letter and letter strings with spoken sounds. With further instruction and experience, they move to a phase where they start to decode and read whole written words. That is, children demonstrate a developmental trend moving from processing small linguistic units (e.g., letters) to larger ones (e.g., rimes, words). As such, it can be hypothesized that sublexical units might have greater impact on word recognition in young children, whereas morphemes and words might be more influential in advanced readers. Another reason that we need databases specialized for children is that children's reading materials differ drastically from those of adults. The properties of words children encounter thus differ from those adults read. For example, the word *ant* might occur quite frequently in children's picture books but not as much in adults' reading materials. Thus, *ant* might be a higher-frequency word in a child database than in an adult one. Indeed, correlations in frequency values between adults' and children's databases have been found to be significantly lower than correlations among adults' databases (van Heuven et al., 2014). In sum, adult corpora cannot sufficiently reflect children's reading experience, and therefore, it is crucial to collect specialized written corpora for children.

To this aim, a number of children's databases have been built for languages with alphabetic orthographies, such as the Children's Printed Word Database (CPWD) (Masterson et al., 2003) and its most recent version in English (Masterson et al., 2010), the childLex-German database in German (Schroeder et al., 2014), the MANULEX in French (Lété et al., 2004), the ONESC database in Spanish (Martínez Martín & García Pérez, 2008), the HelexKids in Greek (Terzopoulos et al., 2017) and the EXCOLEX database in Portuguese (Soares et al., 2014). These systematic counts of the content of children's reading materials provide insights into the properties of written words they are exposed to. For example, CPWD provides word frequencies and orthographic and phonological properties for 12,193 English words that appear in 556 books for children aged 5–9 years. To account for the differences in children's experience with words at different developmental stages, the childLex-German and the MANULEX French databases provide linguistic norms for lexical and sublexical variables by different age groups/grade levels. These databases have been widely used in reading acquisition research and have reduced the time and effort researchers spent constructing experimental stimuli.

However, in Chinese, written word databases for children are limited. Current databases were based on small samples of reading materials, and none was publicly accessible. The most known ones are two analyses of children's textbooks (Shu et al., 2003; Xing et al., 2004). Because their aim was to profile the linguistic features of Chinese characters children are explicitly taught, the materials were limited to several tens of curricular books. Also, neither of these databases calculated important indices like character or word frequencies. For example, the widely cited Shu et al. (2003) study analyzed 12 volumes of textbooks used in Beijing. The total number of character tokens were not reported, but an estimate would be less than 100,000 characters. Xing et al. (2004) also sampled school textbooks and built a corpus of 160,342 character tokens, which also does not suffice to generate valid frequency counts. As noted earlier, word frequency is arguable the most established predictor in word reading. Yet the frequency effect in Chinese children's visual word recognition has been scarcely investigated, perhaps partly due to the lack of a large and representative corpus built with children's reading materials.

More recently, Huang et al. (2020) sampled 52 textbooks used in Jiangsu Province and an additional 43 storybooks. This has led to a corpus of 2.65 million characters and 1.83 million words. The researchers computed frequency and contextual diversity values for characters and words in the corpus. Contextual diversity (CD) is indexed by the number of unique texts/documents in which a word occurs across a corpus (Adelman et al., 2006), which has been suggested to be a better predictor than frequency of lexical processing because it ignores repetition in context (Jones et al., 2017; Perea et al., 2013). Utilizing the frequency and CD indices from the corpus, the researchers have shown in two experiments that grade 4 children's lexical decision times on characters and words were affected by contextual diversity but not frequency.

Although Huang et al.'s (2020) CJC database is much larger compared with the previous corpora and it calculated character and word frequencies, the size of CJC might still be inadequate to provide highly reliable frequency data, which motivated us to build the CCLOWW. The first limitation still has to do with the size of the corpus. In Brysbaert and New's (2009) seminal paper on frequency norms, they addressed the critical question: how large a cost-effect corpus should be. The researchers correlated English word frequencies on various sections of an 88 million-word corpus (500,000 words, 1 million words, … , the complete corpus) with lexical decision times from the English Lexicon Project (Balota et al., 2007). They found that the optimal corpus size depends on the frequency of the words of interest: the frequency counts for high-frequency words became stable once the corpus reached one million words, whereas low-frequency words required a corpus size of at least 16 million words. They thus concluded that a corpus should have 16–30 million words in order to provide reliable frequency norms.

Therefore, the Huang et al. (2020) corpus could be considered small for valid frequency values for the low-frequency characters and words. This could potentially result in the null finding with the frequency effect, particularly given that they used a parametric design which suffers the risk of selection bias. Also note that although there has been some evidence that the English word frequency effect in children's lexical decision was eliminated once contextual diversity was accounted for (Perea et al., 2013), no study has suggested this was also the case in Chinese children except for Huang et al. (2020). This finding thus needs to be replicated before any conclusion can be reached. To do so, we need a large enough developmental corpus that provides reliable frequency counts to devise experimental stimuli. We aim to address this issue in CCLOWW. Additionally, we made CCLOWW freely accessible online to benefit future research in children's reading development.

CCLOWW is a lexical database of simplified Chinese words for children. The materials were extracted from a corpus of 2131 books intended for children in mainland China. We provide separate norms for children of grade 2 and below (beginning readers), grades 3–4 (intermediate readers) and grades 5–6 (experienced readers). In contrast to some previous databases for children (e.g., Shu et al., 2003), we included books children would read in their spare time in addition to school textbooks. This is because children's extracurricular reading activities contribute strongly to their print exposure and reading development (e.g., Mol & Bus, 2011). There is even evidence that time spent on reading outside school is the best predictor of progress of reading achievement in school (Cunningham & Stanovich, 1997). Also, textbooks used across mainland China are highly heterogeneous, making it difficult to gauge the typical children's reading materials solely based on them.

In CCLOWW, we provide statistics for characters in addition to words. This is because current research suggests that character reading might be different from word reading. A character is the smallest reading unit in Chinese, corresponding to a spoken syllable and, in many cases, a morpheme. Most of them can themselves stand alone as meaningful, independent words (e.g., 花, flower). However, multicharacter words are and much more common (e.g., 花生, peanut). According to Tan and Perfetti's (1999) analysis of the Modern Chinese Frequency Dictionary (Wang et al., 1986), Chinese words contain from one to four characters and more than 70% of them comprise of two or more characters. Both characters and words affect reading in Chinese. It has been reported that the recognition time (Zhang & Peng, 1992) and lexical decision time (Cai & Brysbaert, 2010) of two-character words were affected by the frequency of not only the words but also the composing characters. Studies using eye-movement tracking have also demonstrated differential effects of character and word properties on word processing during sentence reading (Bai et al., 2008; Chen et al., 2003; Chen & Ko, 2011; Yan et al., 2006). These findings provide evidence that character reading and word reading are different, and they should be treated separately in psycholinguistic research. Therefore, there is a strong motivation that a comprehensive child database should analyze characters and words separately and provide information for both.

In the following, we first describe our approaches to corpus collection and linguistic processing. Next, we provide detailed information of the length, syntactic categories, frequency and contextual diversity statistics of the characters and words with a particular focus on grade-related changes. We then ran correlational analyses among the frequency indices computed in the CCLOWW subcorpora. They were also compared with two other Chinese frequency databases: CJC (Huang et al., 2020) and SUBTLEX-CH (Cai & Brysbaert, 2010). We chose these two databases because the former was the only corpus that computed frequency statistics based on children's reading materials in Chinese. The latter was built from a large sample of movie subtitles that has been shown to provide highly reliable frequency and contextual diversity data for simplified Chinese. We expected the frequency measures of CCLOWW to correlate well with that of the two databases. Finally, to validate that CCLOWW is a reliable database for Chinese character and word frequency for children and that it might also explain some variance in adults' written word processing, we ran a series of analyses on written word recognition data we collected or acquired elsewhere. First, we regressed character-naming accuracy data from grade 2–3 children on the frequency and contextual diversity measures from CCLOWW and from CJC. We expected that CCLOWW would reliably predict the children' character-naming accuracy and that it might outperform CJC. Second, we conducted a word-naming experiment with grade 2–3 children, in which we manipulated the frequency of one- and two-character words based on measures from CCLOWW and from SUBTLEX-CH (Cai & Brysbaert, 2010). The purpose was to see whether we could find a frequency effect on children's word reading based on data from CCLOWW and from the adult database. Lastly, we used word frequency statistics from CCLOWW and from SUBTLEX-CH to predict adults' word-naming and lexical decision reaction times (RTs). The naming data were collected with a convenient sample and the lexical decision data was obtained from a published megastudy (Tsang et al., 2018). We expected that the frequency measures from both databases could predict adults' written word recognition latencies and that CCLOWW might explain extra variance in addition to the adult database. This result would indicate that early print exposure might have long-term influences on readers' visual word processing.

Method

Corpus collection

We sampled curricular books published by the People's Education Press and popular extracurricular books. To ensure that the corpus reflects children's average reading experience in mainland China, we selected the extracurricular books using the following criteria. First, we obtained the 2019 Book Recommendation for Primary and Middle School Libraries (Ministry of Education, 2019) and the 2020 Extracurricular Reading Recommendation for Primary and Middle School Students (Ministry of Education, 2020), provided by the Ministry of Education of the People's Republic of China (henceforth referred to as the MOE lists). These lists included several thousands of books arranged by recommended grade level: grades 1–2, grades 3–4 and grades 5–6. We also followed this grade level classification to build the current corpus, because most of the books we selected were obtained from these lists. Next, we analyzed the 2019 sales figures of children's books provided by the two most popular online bookstores in China: www.book.jd.com (Jingdong) and www.book.dangdang.com (Dangdang). We selected books with sales ranking between 1 and 100 by three age groups as provided by the websites: 4–6 years, 7–10 years and 11–14 years. Given the increasing popularity of electronic books among Chinese children (National Press and Publication Administration, 2020), we also sampled from an online reading website (www.xiaoshuotxt.com) and obtained a list of children's literature ranking in popularity from 1 to 100. Recommended reading levels were not provided in this list. Finally, given that we could not obtain as many books intended for children at or below grade 2 from the above sources, we decided to acquire additional picture books to ensure there were enough characters/words in this subcorpus. It should be noted that these criteria were not applied rigidly but were used as an orientation for selecting book materials. If a book was not available to us, we simply replaced it with another one from the same list.

In this corpus, books were assigned to three grade levels: grade 2 and below (G2), grades 3–4 (G34) and grades 5–6 (G56). Because reading level recommendation was not available for some of the materials (either from the MOE list or from publishers), we used the following procedure to assign the books to the three grade levels: First, all picture books were assigned to the G2 subcorpus, as they are intended for shared book reading with beginner readers. Second, if a book was on the MOE lists, we followed the grade level recommendation of the lists. Next, when a recommended age was available (as provided by the online bookstores), we converted the age to corresponding grade level: 4–6 to G2; 7–9 to G34; 11–14 to G56. Finally, when none of the criteria was applicable for a book, we asked two primary school teachers of Chinese language to recommend a reading level and resorted to a third teacher in case of disagreement. If all three teachers responded that they had not read or heard of a book, then this book was excluded from the corpus.

Within each subcorpus, very long books (G2: 2; G34: 5; G56: 5) with a character count greater than 2.5 standard deviations from the subcorpus' mean were subdivided into no more than five shorter documents. This led to a total of 2,152 unique documents. Table 1 presents the numbers of documents in each grade level per subject area. Picture books in G2 were all categorized as literature.

Linguistic processing and cleaning

The printed books were photographed and saved as PDF files. All pages were included, except for headers, prefaces, introductory notes, bibliography, references and advertisements. They were converted into UTF-8 text files using the HUDUN character recognition system (www.huduntech.com) or Baidu OCR (https://ai.baidu.com/tech/ocr). The texts were proofread and corrected by the first author and two research assistants.

Word segmentation and POS tagging were conducted using fastHan (Geng et al., 2020). fastHan is a BERT (Bidirectional Encoder Representations from Transformers)-based Chinese natural language processing toolkit trained and evaluated on 13 corpora. It has shown high accuracy on Chinese word segmentation and POS tagging, outperforming other popular Chinese segmentation tools such as Jieba and SnowNLP (Geng et al., 2020). The output of fastHan provided the lines of the words and the part of speech. Following Cai and Brysbaert (2010), we cleaned the output files by removing noncharacters, such as English letters and Arabic numerals, found in low-frequency sequences. We then used

Table 1 Numbers of documents in each grade level by subject area.

	All	G2	G34	G56
Literature	1784	1532	151	100
Education	80	4	51	21
History and culture	67	0	46	22
Science and technology	47	5	21	21
Astronomy and geography	49	1	19	29
Mathematics	27	4	18	5
Biology	36	4	18	14
Language	31	5	18	8
Physiology and health	10	0	7	3
Arts	25	7	10	12
Total	2152	1558	359	235

the Table of General Standard Chinese Characters (Ministry of Education, 2013) as a reference to remove characters that are no longer used in contemporary Chinese. Additionally, we removed the following items which we decided were nonwords: (1) items with a length greater than 15 characters and a frequency count less than 2; (2) non-nouns with a word length greater than eight characters; (3) interjections with a length greater than four characters; (4) number words with a frequency count less than 10. An example of items removed from these procedures is "啦啦啦啦啦啦啦啦啦", which was a repetition of the interjection "啦".

Frequency calculation

Frequency measures were calculated based on the total counts and on the number of documents the words occurred in (i.e., contextual diversity). In addition to raw counts and frequency per million, we computed *Zipf* value, which is a standardized frequency index expressed on a logarithmic scale. We used the formula from van Heuven et al. (2014) to compute this index: log10 (counts per million) + 3. The *Zipf* value ranges from 1 to 7 with a *Zipf* = 1 (1 per 100 million) suggesting very low frequency; a *Zipf* = 6 (1 per 1000) representing very high frequency; and a *Zipf* between 3 (1 per million) and 4 (1 per 100,000) suggesting medium frequency. An advantage of this scale is that it allows researchers to select items with a frequency below 1 per million that otherwise would have been excluded (van Heuven et al., 2014). We also computed a log-transformed contextual diversity value (logCD).

Results

Summary of the corpus

In this article, data reported and used in the experiments are based on calculations of lemma entries, except for distributions of word syntactic categories. Data for non-lemmatized POS-tagged words are also available in the online database.

Table 2 presents a summary of the corpus and the subcorpora. The final corpus comprises 2,152 documents: 1558 in the G2 subcorpora (2,286,344 character and 1,554,243 word tokens), 359 in the G34 subcorpora (16,181,732 character and 10,542,080 word tokens) and 235 in the G56 subcorpora (16,203,348 character and 10,393,810 word tokens). Thus, the character counts of the total corpus and the G34 and the G56 subcorpora have met the criteria set by Brysbaert and New (2009) for being able to provide reliable frequency norms. On average, the document for beginning readers contains approximately 1467 characters, 998 words; the document for intermediate readers 45,074 characters, 29,365 words; and the document for advanced readers 68,950 characters, 44,229 words. The significant increase in the average number of characters and words per document indicates the advancement of children's reading ability.

The distributions of word frequency in CCLOWW demonstrated some patterns different from child corpora in other languages. Words that occur only once (hapax words) account for a fair amount of the corpus, particularly of the G2 subcorpora, but their occurrence accounts for only 0.75%, 0.19% and 0.17% of all word tokens in the subcorpora, respectively. By contrast, less than half of the word types occur more than five times, but their occurrence accounts for 99.25% of all tokens (G2: 97.29%; G34: 98.79%; G56: 98.68%). The 500 most frequent characters account for 58.85% of all tokens. Although a high percentage of low-frequency words has also been found in other child corpora, hapax words are less common in CCLOWW. There are no hapax words in the total corpus, and the proportions of hapax words in the G34 and the G56 subcorpora are lower than 20%. In comparison, hapax words can constitute up to 49% of the word types in the other child corpora (e.g., Schroeder et al., 2014). This difference may be because in deep orthographies, children need more exposure to acquire

Table 2 Distribution of character and word tokens, types, grade-specific items and words occurring once and more than five times in the corpus and the subcorpora

	Corpus			
	All	G2	G34	G56
Character tokens	34,671,424	2,286,344	16,181,732	16,203,348
Word tokens	22,427,010	1,554,243	10,542,080	10,393,810
Character types	6746	4351	6285	6406
Word types	153,079	37,516	120,416	125,459
% grade-unique characters (type)	NA	1.84	4.96	6.82
% grade-unique words (type)	NA	5.02	19.69	23.28
% words (type) occurring once	0	32.31	17.45	16.12
% words (type) occurring five or more times	55.79	36.91	48.14	47.30

words than in shallow orthographies. There is evidence that while a single exposure might suffice for word acquisition in pointed Hebrew, a shallow script (Share, 2004), children typically need to see an English word four or more times to be able to recognize it fluently (Bowey & Muller, 2005). In Chinese, children might need even more exposure for necessary learning, and hence the smaller number of hapax words found in CCLOWW.

Some developmental trends of children's experience with written Chinese are hinted at by cross-subcorpus comparisons. Table 2 shows that the percentages of subcorpus-unique characters and words increase with grade level. Additionally, G34 subcorpus contains a larger proportion of novel characters but a smaller proportion of novel words than G56 subcorpus. Characters in G34 that are not in G2 occupy 31.15% of all G34 character tokens, whereas G56 characters that do not occur in G34 occupy only 7.07% of G56. In comparison, words that occur in G34 but not in G2 account for 71.72% of all G34 words, while words that occur in G56 but not in G34 account for 24.53% of all G56. These results show that children's printed vocabulary grows substantially throughout the stages of reading development. The characters in the G2 subcorpus constitute the basic reading vocabulary for beginning readers, which increases dramatically when they move from beginner to intermediate reader status but not as much when they advance further. That is, Chinese children may have already encountered most of the characters around grades 3 and 4. This finding is consistent with a previous analysis of school textbooks, which suggested that the task of learning characters is the heaviest for second and third grade children and the number of new characters introduced at school decreases in higher grades (Shu et al., 2003). In comparison, the proportion of grade-specific words increases across the subcorpora, indicating that children keep encountering new words as their reading experience grows.

Word length

In the corpus, 4527 of the words are single-character words (G2: 2600; G34: 4282; G56: 4341); 89,849 are two-character words (G2: 26,147; G34: 74,916; G56: 76,386); 40,451 are three-character words (G2: 5896; G34: 28,253; G56: 29,618); and 14,737 are four-character words (G2: 2736; G34: 11,181; G56: 12,424). Distribution of the word length in each subcorpus is presented in Table 3. The result is consistent with previous analysis of modern Chinese dictionaries. That is, more than 70% Chinese words comprise two or more characters (Tan & Perfetti, 1999). Although according to the database, two-character words are the most typical type, one-character words should be the most seen by children. There is a decreasing trend in the percentage of one-character words and an increasing trend in that of words with more than two characters across grades. This indicates that as children's reading experience accumulates, they see and acquire more complex words.

Syntactic categories

Table 4 shows the percentages of words by syntactic categories across the grade levels. fastHan uses the Penn Chinese Treebank label sets (Xue et al., 2019) for part-of-speech (POS) tagging, which contains 33 labels. In our analysis, some of the labels were combined for convenience. For example, common nouns, proper nouns and temporal nouns were all treated as nouns. Content words (nouns, verbs, adjectives, adverbs, numerals and measure words) account for over 90% of all word types. Nevertheless, their occurrence is less frequent, taking up just over 50% across the subcorpora. The proportion of nouns increases steadily from 22.65% to 27.17% across the subcorpora. This is possibly because the documents in the G34 and G56 subcorpora covered a wider range of themes than in the G2 subcorpus, which may contain many new concepts and terms. In comparison, the proportions of other content words decrease from the G2 to the G56 subcorpora.

Frequencies

Figure 1 displays the distributions of frequency (*Zipf*) and contextual diversity (logCD) of characters and words by subcorpus in CCLOWW. Dashed lines indicate the 10%, 25%, 50%, 75% and 90% percentiles. The *Zipf* values of characters and words across the subcorpora range from 1.27 to 7.62, with means ranging from 2.65 to 4.31. In particular, the distributions of word frequencies is consistent with the *Zipf* law (van

Table 3 Distribution of word length in the corpus and the subcorpora

	All		G2		G34		G56	
	Type	Token	Type	Token	Type	Token	Type	Token
% One-character words	2.96	54.00	6.93	58.41	3.56	54.24	3.46	52.94
% Two-character words	58.69	41.17	69.70	37.84	62.21	41.01	60.89	41.97
% Three-character words	26.42	3.59	15.72	2.95	23.46	3.59	23.61	3.68
% Four-character words	9.63	1.12	7.29	0.76	9.29	1.05	9.90	1.26

Table 4 Distribution of syntactic categories of words in the corpus

	Word types				Word tokens			
	All	G2	G34	G56	All	G2	G34	G56
% Noun	63.32	49.05	58.86	59.22	26.56	22.65	26.03	27.17
% Verb	23.22	31.11	24.86	24.62	23.10	24.14	23.42	22.35
% Adjective	8.27	10.12	9.68	9.80	5.19	5.88	5.15	5.22
% Adverb	2.88	4.78	3.54	3.47	12.01	12.48	12.38	11.66
% Preposition	0.16	0.39	0.21	0.20	3.44	2.93	3.31	3.63
% Pronoun	0.16	0.29	0.20	0.17	6.12	7.95	6.20	5.82
% Numeral	0.22	0.62	0.44	0.41	2.85	2.90	3.01	3.08
% Particle	0.10	0.22	0.16	0.13	2.03	2.99	2.12	1.81
% Conjunction	0.10	0.24	0.13	0.13	1.86	1.70	1.82	1.93
% Measure word	0.52	0.95	0.65	0.62	3.35	3.30	3.32	3.43
% Determiner	0.008	0.02	0.01	0.01	1.73	1.45	1.71	1.80
% Interjection	0.39	0.88	0.36	0.36	0.17	0.50	0.16	0.12
% Onomatopoeia	0.31	0.72	0.29	0.28	0.05	0.16	0.05	0.04
% Other	0.34	0.61	0.61	0.58	11.54	10.97	11.32	11.94

Heuven et al., 2014) with means lying around 3 and 4. There is also a decreasing trend of mean character (G2: 4.31; G34: 3.95; G56: 3.92) and word frequencies (G2: 3.44; G34: 2.82; G56: 2.80) from lower to higher grades, which might indicate that children see the same words more repeatedly at the beginning stages of reading development than they do later on.

The mean overall word frequency is low in all grade levels. Words that occur up to times account for about 50% of the total corpus. On the other hand, high-frequency words, normally considered as those occurring 100 or more times per million words, are rather rare (All = 0.7%; G2 = 3.72%; G34 = 1.27%; G56 = 1.18%). This bias towards low-frequency words is also demonstrated in the contextual diversity (CD) index. On average, CD varies between 1 and 2128 across subcorpora with a 50% percentile of four documents in the total corpus. That is, most of the words are context-specific. These findings are in line with the finding of large numbers of hapax words in each subcorpora. This shows that, like other child databases (Lété et al., 2004; Soares et al., 2014; Terzopoulos et al., 2017), CCLOWW has a strong bias toward low-frequency words, which is expected by the *Zipf* law (Zipf, 2016) in constructing corpora. The average *Zipf* values and logCD for both characters and words decline across subcorpora, suggesting that characters and words occur more repeatedly in lower than in upper grade levels.

Correlations among subcorpora and with existing databases

Correlations amongst CCLOWW and the subcorpora on frequency and contextual diversity measures were all significant and strong (Table 5). The coefficients for characters were higher than that for words.

We also conducted correlation analyses on character and word frequency and CD values computed from our corpus and from CJC (Huang et al., 2020) and SUBTLEX-CH (Cai & Brysbaert, 2010). CJC was the only corpus that computed frequency statistics based on children's reading materials in Chinese. SUBTLEX-CH has been shown to provide highly reliable frequency and CD data for simplified Chinese for adults. Note that we were only able to acquire data arranged by grades 1–4 from CJC and not the total corpus. We selected the grade 3 dataset to include in the analysis because it is a middle grade level and should be a closer reflection of the average of the total corpus.

There were 4083 characters and 21,881 words shared across CCLOWW, CJC and SUBTLEX-CH. Table 5 presents the results of the correlational analyses between the measures from the three databases for the common characters and words. The correlations of character and word frequency and contextual diversity are reasonably high. The correlations between CCLOWW and SUBTLEX-CH were higher than that between CCLOWW and CJC or between CJC and SUBTLEX-CH. This is possibly because our corpus and SUBTLEX-CH are both based on large character and word counts. Yet interestingly, the G2 subcorpus correlated better with CJC, whereas the G56 subcorpus correlated better with SUBTLEX-CH, possibly reflecting that the materials in the G56 subcorpus were closer to adults' reading materials.

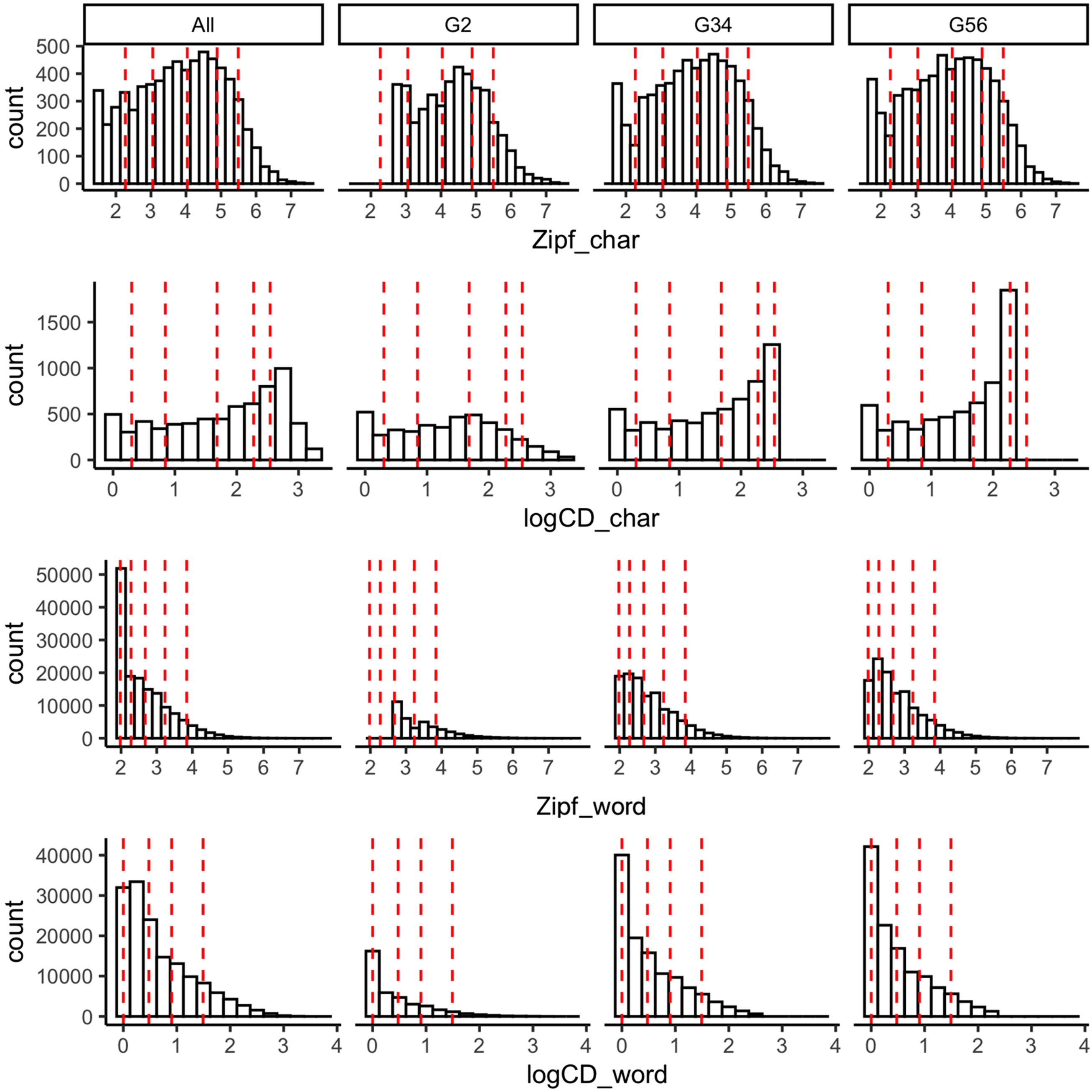

Fig. 1 Distributions of character and word frequency (Zipf) and contextual diversity (logCD) in the corpus and subcorpora. Dashed vertical lines indicate the 10%, 25%, 50%, 75% and 90% percentiles. Note. The 10% and 25% percentiles of logCD_word are both at 0. Zipf_char = character Zipf; logCD_char = character log-transformed CD; Zipf_word = word Zipf; logCD_word = word log-transformed CD

Validations of CCLOWW frequencies

Predicting grade 2–3 children's character naming

We analyzed character-naming accuracy data collected from 52 7–9-year-old children (20 males, age *mean* = 8.33, *SD* = 0.31, range = 7.78 to 9.05). The children, and the participants in the following word-naming experiment, were recruited as part of a large neuroimaging study. The study was approved by the East China Normal University Committee on Human Research Protection. Written consent was obtained from the parents. The children 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

Table 5 Pearson's correlations between CCLOWW and the subcorpora, CJC grade 3 and SUBTLEX-CH

Measure	1	2	3	4	5	6	7	8	9	10	11
Character ($N = 4083$)											
1. *Zipf*_CCLOWW_G2	--										
2. *Zipf*_CCLOWW_G34	0.90	--									
3. *Zipf*_CCLOWW_G56	0.87	0.97	--								
4. *Zipf*_CCLOWW_all	0.90	0.99	0.99	--							
5. LogFreq_CJC Grade 3	0.80	0.77	0.76	0.78	--						
6. LogFreq_SUBTLEX-CH	0.82	0.90	0.90	0.90	0.70	--					
7. LogCD_CCLOWW_G2	0.97	0.89	0.87	0.90	0.79	0.83	--				
8. LogCD_CCLOWW_G34	0.80	0.91	0.89	0.90	0.67	0.85	0.85	--			
9. LogCD_CCLOWW_G56	0.75	0.87	0.89	0.88	0.64	0.83	0.80	0.98	--		
10. LogCD_CCLOWW_all	0.87	0.94	0.93	0.94	0.72	0.88	0.91	0.98	0.97	--	
11. LogCD_CJC Grade 3	0.92	0.91	0.90	0.91	0.83	0.85	0.94	0.86	0.82	0.90	--
12. LogCD_SUBTLEX-CH	0.78	0.86	0.86	0.86	0.66	0.97	0.81	0.88	0.87	0.89	0.83
Word ($N = 21{,}881$)											
1. *Zipf*_CCLOWW_G2	--										
2. *Zipf*_CCLOWW_G34	0.78	--									
3. *Zipf*_CCLOWW_G56	0.73	0.92	--								
4. *Zipf*_CCLOWW_all	0.80	0.98	0.97	--							
5. LogFreq_CJC Grade 3	0.81	0.73	0.70	0.73	--						
6. LogFreq_SUBTLEX-CH	0.62	0.71	0.71	0.72	0.64	--					
7. LogCD_CCLOWW_G2	0.96	0.79	0.74	0.80	0.80	0.63	--				
8. LogCD_CCLOWW_G34	0.70	0.94	0.89	0.92	0.65	0.67	0.76	--			
9. LogCD_CCLOWW_G56	0.65	0.88	0.93	0.91	0.62	0.67	0.70	0.94	--		
10. LogCD_CCLOWW_all	0.74	0.93	0.93	0.94	0.66	0.69	0.80	0.98	0.97	--	
11. LogCD_CJC Grade 3	0.81	0.75	0.72	0.75	0.96	0.66	0.82	0.71	0.67	0.72	--
12. LogCD_SUBTLEX-CH	0.59	0.70	0.71	0.72	0.62	0.99	0.62	0.69	0.69	0.71	0.65

ps < .001

just skip it. The task took about 10 minutes. In the experiment, the children named 120 characters, arranged in an increasing order of difficulty, of which 94 are covered by the CCLOWW corpus. We therefore analyzed the naming accuracy for the 94 characters.

Several logistic regression models were built to examine the effects of CCLOWW frequency and CD on the children's character-naming accuracy (binarily coded as 1 = correct, 0 = incorrect). To avoid the issue of multicollinearity, frequency and CD were analyzed in separate models. In Models 1 and 2, the character frequency (*Zipf*) and character log-transformed CD (logCD) were from the total corpus; in Models 3 and 4, they were from the G34 subcorpora. We also obtained age of acquisition (AoA), regularity, number of strokes, concreteness and homophonic neighborhood density from Liu et al. (2007), and added them in the models as predictors. These are some of the variables that are known to affect character reading in Chinese adults (e.g., Liu et al., 2007). The results are shown in Table 6. Note that in all analyses in this article, the variance inflation factors (VIFs) were all smaller than 2.65, indicating that there was no issue of multicollinearity.

All models provided a good fit to the children's naming data, as indicated by McFadden's pseudo R^2. A rule of thumb is that McFadden's pseudo R^2 between 0.2 and 0.4 indicates excellent model fit (McFadden, 1977). In both models, character frequency (*Zipf*) and contextual diversity (logCD) reliably predicted the children's naming accuracy of the characters. Consistent with research on character recognition with adults (Chen et al., 2009; Liu et al., 2007; Sze et al., 2015), AoA also reliably predicted naming accuracy in the children. Character visual complexity (number of strokes) significantly predicts naming in all models but Model 1.

To compare the predictability of CCLOWW character frequencies with that of CJC, the other written word corpus for Chinese children, we additionally constructed Models 5 and 6, in which CJC character frequency and CD were predictors in respective models. All other predictors remained the same. The effects of CJC character frequency and CD were also significant (see Table 6). However, compared with

Table 6 Results of logistic regression analysis for variables predicting children's character-naming accuracy

Predictor	*B*	*SE*	*z*	*p*	*OR*
Model 1					
*Zipf*_CCLOWW all	2.01	0.15	13.71	< .001	7.49
Regularity	0.35	0.05	7.27	< .001	1.41
AoA	−1.39	0.10	−14.21	< .001	0.25
McFadden's R^2 = .283					
Model 2					
LogCD_CCLOWW all	3.49	0.27	12.70	< .001	32.70
Regularity	0.39	0.05	8.07	< .001	1.48
AoA	−1.50	0.10	−15.34	< .001	0.22
N strokes	0.03	0.01	2.06	.040	1.03
McFadden's R^2 = .279					
Model 3					
*Zipf*_CCLOWW G34	1.82	0.15	12.20	< .001	6.19
Regularity	0.32	0.05	6.82	< .001	1.38
AoA	−1.42	0.10	−13.81	< .001	0.24
N strokes	0.03	0.01	2.34	.019	1.03
McFadden's R^2 = .277					
Model 4					
LogCD_CCLOWW all	3.16	0.27	11.63	< .001	23.53
Regularity	0.38	0.05	7.79	< .001	1.46
AoA	−1.61	0.10	−16.72	< .001	0.20
N strokes	0.03	0.01	2.31	.021	1.03
Concreteness	−0.10	0.05	−1.99	.046	0.91
McFadden's R^2 = .274					
Model 5					
LogFreq_CJC	0.84	0.08	10.55	.008	2.31
Regularity	0.30	0.05	6.40	< .001	1.35
AoA	−1.95	0.10	−20.52	< .001	0.14
N strokes	0.04	0.01	2.81	.005	1.04
Concreteness	−0.36	0.04	−8.41	< .001	0.70
Homophonic neighborhood	0.00	0.00	−2.05	.004	1.00
McFadden's R^2 = .277					
Model 6					
LogCD_CJC	1.52	0.12	12.47	< .001	4.59
Regularity	0.31	0.05	6.58	< .001	1.37
AoA	−1.66	0.10	−16.58	< .001	0.19
N strokes	0.03	0.01	2.01	.045	1.03
Concreteness	−0.31	0.04	−7.17	< .001	0.73
Homophonic neighborhood	0.00	0.00	−2.26	.024	1.00
McFadden's R^2 = .285					

In Models 1 and 2, the frequency and contextual diversity measures were from the total corpus of CCLOWW; in Models 3 and 4, the measures were from the G34 subcorpus; in Models 5 and 6, the measures were from the CJC Grade 3 subcorpus. All other predictors were the same in the three models. For brevity, only significant predictors in each model are presented

Models 1 and 3, where the odds ratios (*OR*) of character frequency were 7.49 and 6.19 respectively, the *OR* of frequency in Model 5 was only 2.31. Similarly, the *OR* of CD in Models 2 and 4 were much higher than that in Model 6. This finding indicates that compared with CJC, CCLOWW character frequency and CD measures might have greater impacts on children's character-naming accuracy.

Predicting grade 2–3 children's word naming

Our next validation was to assess the effect of word frequency computed from CCLOWW on children's word-naming accuracy. Additionally, we wanted to see whether word frequencies computed in adult corpora also affect young children's word naming. To these aims, we collected naming accuracy data on 160 Chinese words (80 one-character words and 80 two-character words) from 33 grade 2–3 children (19 males, age *mean* = 8.43, *SD* = .35, range = 7.92 to 9.15). The procedure was identical to the aforementioned character-naming task.

The stimuli words were selected such that four experimental conditions were created differing on high and low frequency from CCLOWW (child corpus frequency: high vs. low) and SUBTLEX-CH (adult corpus frequency: high vs. low). Distribution of the sampled words' log-transformed frequencies are presented in Fig. 2. The high and low conditions differed significantly on word frequencies indexed from CCLOWW (one-character words: $t(359.07) = 59.22$, $p < .001$; two-character words: $t(436.75) = 8.37$, $p < .001$) and from SUBTLEX-CH (one-character words: $t(1621.3) = 177.98$, $p < .001$; two-character words: $t(436.75) = 8.37$, $p < .001$). Words in the four conditions were also matched on AoA, number of strokes and concreteness (see Table 7) for one- and two-character words separately. Statistics of the controlled variables were obtained from Liu et al. (2007) for the one-character words and from Xu and colleagues (Xu et al., 2021; Xu & Li, 2020) for the two-character words.

The mean naming accuracies per condition are shown in Table 7. We constructed two separate logistic mixed-effects regression models for one- and two-character words. Naming accuracy (binarily coded as 1= correct, 0 = incorrect) was the dependent variable. Child corpus frequency (high vs. low), adult corpus frequency (high vs. low) and their interaction were the fixed effects. Following Barr et al.'s (2013) suggestion, models were initially constructed with a maximal random structure with by-subject and by-item intercepts and slopes. Random structures that explained the smallest variance were then removed one at a time to facilitate model convergence. The final converged model structure was as follows: response ~ child*adult + (1 | subject) + (1 | item). The significance of the fixed effects was tested with a likelihood ratio test, comparing a fuller model to a reduced model with one fewer fixed effect.

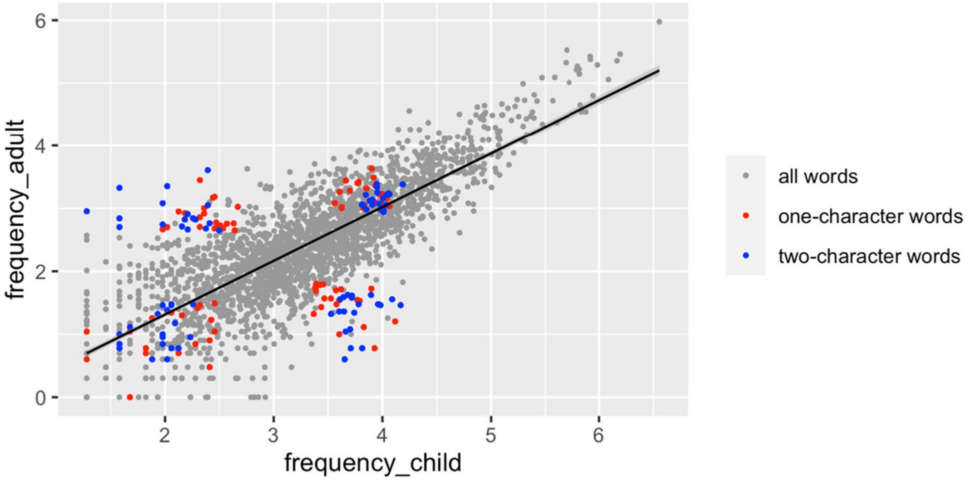

Fig. 2 Frequency distributions of the experimental items. Frequency_child is the Zipf value in the total corpus of CCLOWW. Frequency_adult is the log-transformed word counts in SUBTLEX-CH. Grey dots represent the common two-character words in the two databases. Colored dots represent the sampled words in the children's word-naming experiment

Table 7 Means of descriptive data of the experimental words in each condition (standard deviations in parentheses)

	HcHa	HcLa	LcHa	LcLa
One-character words (*N* = 80)				
*Zipf*_CLOWW	4.50 (.16)	4.28 (.21)	3.06 (.20)	2.70 (.37)
LogFreq_SUBTLEX-CH	3.23 (.19)	1.50 (.29)	2.86 (.21)	0.97 (.38)
N strokes	10.25 (1.97)	10.30 (4.03)	9.40 (2.48)	8.95 (3.30)
AoA	4.42 (.18)	4.69 (.68)	4.61 (.71)	4.70 (.31)
Concreteness	4.77 (.71)	5.12 (.78)	4.48 (1.37)	4.81 (1.07)
Regularity (*N* regular/irregular/non-phonogram)	2/5/13	6/6/8	5/9/9	7/6/7
Naming accuracy	.85 (.36)	.62 (.49)	.76 (.43)	.49 (.50)
Two-character words (*N* = 80)				
*Zipf*_CLOWW	4.63 (.09)	4.43 (.17)	2.77 (.34)	2.62 (.20)
LogFreq_SUBTLEX-CH	3.14 (.14)	1.33 (.31)	2.90 (.26)	1.04 (.30)
N strokes	16.20 (4.46)	15.25 (4.40)	16.80 (4.82)	16.55 (4.35)
AoA	12.77 (1.29)	11.98 (1.03)	12.62 (2.95)	12.05 (1.31)
Concreteness	2.42 (.35)	2.32 (.70)	2.52 (.71)	2.64 (.54)
Naming accuracy	.85 (.36)	.78 (.41)	.74 (.44)	.69 (.46)

HcHa high frequency in child and high frequency in adult corpora, *HcLa* high frequency in child and low frequency in adult corpora, *LcHa* low frequency in child and high frequency in adult corpora, *LcLa* low frequency in child and low frequency in adult corpora. Zipf_CLOWW is the zipf value in the total corpus of CCLOWW. LogFreq_SUBTLEX-CH is the log-transformed word counts in SUBTLEX-CH

For both one- and two-character word naming, there was no significant interaction effect (one-character words: $\chi2 = .28$, $df = 1$, $p = .597$; two-character words: $\chi2 = .09$, $df = 1$, $p = .763$). For one-character word naming, the effect of frequency indexed from both the child and the adult corpora was significant (child: $\chi2 = 5.14$, $df = 1$, $p = .023$; adult: $\chi2 = 16.27$, $df = 1$, $p < .001$). However, for two-character word naming, only the frequency indexed from the child corpus was significant, $\chi2 = 6.91$, $df = 1$, $p = .009$. The effect of frequency from the adult corpus was not significant, $\chi2 = 2.79$, $df = 1$, $p = .09$. The results show a clear effect of word frequency based on CCLOWW on the children's one- and two-character word-naming accuracy.

Predicting adults' word naming and lexical decision

To see whether CCLOWW frequencies could also predict adults' written word recognition, we ran an additional naming experiment with a convenience sample of 22 Chinese-speaking adults (13 males, age *mean* = 25.76, range = 23–30). All were native Chinese speakers. We sampled 279 two-character words that did not correlate much on CCLOWW and SUBTLEX-CH frequency measures as target words (Fig. 3). In the experiment, participants 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 2000 ms. Participants were tested individually in a quiet room at East China Normal University. They were asked to say the word aloud as quickly and accurately as they could. Optional breaks were provided after every 93 trials. The task took about 15 minutes. The study was approved by the East China Normal University Committee on Human Research Protection. Signed written consent was obtained from all participants.

For each participant, speech errors, incorrect naming responses and responses with RTs beyond 2.5 standard deviations from the mean were removed (4.04% of the data). The final analysis included 5890 valid observations. Because we wanted to compare the results of the naming experiment with an existing lexical decision dataset, which provided standardized RTs across participants, we also standardized the naming RTs by participant and used the *z* scores in the following analysis.

A linear regression model was built to regress the naming RTs on word frequency in our database (*zipf*_CCLOWW) and word frequency in the adult database (logFreq_SUBTLEX-CH). The following lexical and sublexical variables were also included in the model in order to explain the most variance in the naming RTs: AoA, total number of strokes (*N* strokes), concreteness, frequency of the first character in CCLOWW (*zipf*_C1), frequency of the second character in CCLOWW (*zipf*_C2), stroke counts of the first character (*N* strokes_C1), stroke counts of the second character (*N* strokes_C2), 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), regularity of the first character (Regularity_C1). Statistics of these variables were obtained from Xu et al. (Xu et al., 2021; Xu & Li, 2020) and Sun et al. (2018). The results of the regression model are presented in Table 8. The significant predictors jointly explained 39.58% of the variance in the naming reaction times. Both SUBTLEX-CH and CCLOWW word frequencies significantly predicted the naming RTs. Frequency, stroke counts and regularity of the first character also significantly influenced the naming RTs. The effects of the other well-established predictors of visual word recognition were all significant or approaching significance.

Given that there were discrepancies in the frequency effect on the two most used visual word recognition tasks: word naming and lexical decision (Morrison & Ellis, 1995), we went further to see whether the results would be the same on adults' lexical decision performance. We obtained adults' lexical decision RTs from a megastudy (Tsang et al., 2018) for 278 of the words used in the naming experiment. The same predictors were added in the regression model and the results were largely the same as the word naming experiment (Table 8). Together, the significant predictors explained 57.43% of the variance in the lexical decision RTs.

Finally, we explored whether CCLOWW word and character frequency statistics could explain extra variance in

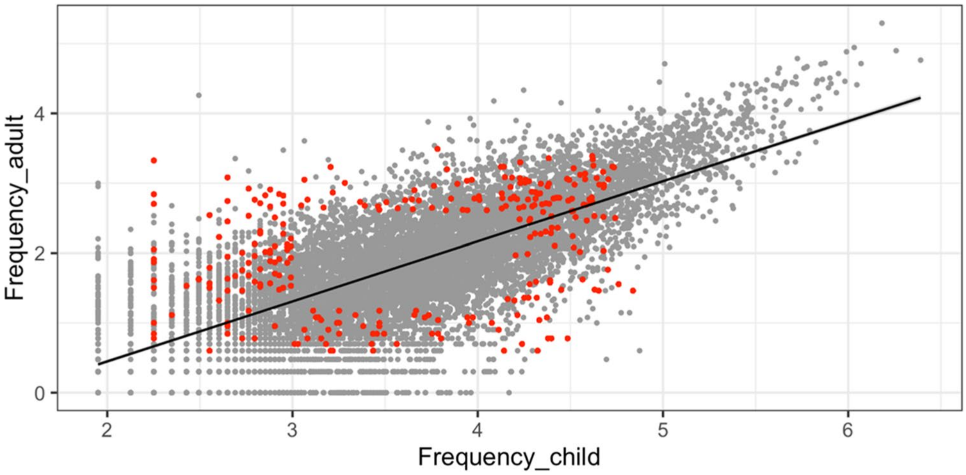

Fig. 3 Frequency distributions of the experimental items. Frequency_child is the Zipf value in the total corpus of CCLOWW. Frequency_adult is the log-transformed word counts in SUBTLEX-CH. Grey dots represent the common two-character words in the two databases. Red dots represent the sampled words in the adults' word-naming experiment

Table 8 Results of linear regression analysis for variables predicting adults' word-naming and lexical decision RTs

Predictor	*β*	*SE*	*t*	*p*
Word naming				
*Zipf*_CCLOWW	−0.05	0.01	−5.18	< .001
LogFreq_SUBTLEX-CH	−0.03	0.01	−3.15	.002
Zipf _C1	−0.26	0.01	−19.34	< .001
N strokes_C1	0.01	0.00	6.11	< .001
Ortho *N*_C1	0.00	0.00	6.75	< .001
Ortho *N*_C2	0.00	0.00	5.77	< .001
SR ortho *N*_C1	0.00	0.00	−9.32	< .001
SR ortho *N*_C2	0.00	0.00	5.34	< .001
PR ortho *N*_C1	0.01	0.00	4.03	< .001
PR ortho *N*_C2	−0.01	0.00	−6.45	< .001
AoA	0.02	0.00	5.38	< .001
Concreteness	−0.02	0.01	−1.86	.064
Regularity_C1	−0.07	0.01	−10.12	< .001
Adjusted R^2 = .396				
Lexical decision				
*Zipf*_CCLOWW	−0.17	0.01	−22.96	< .001
LogFreq_SUBTLEX-CH	−0.18	0.01	−24.96	< .001
Zipf _C2	0.19	0.01	14.64	< .001
N strokes	0.02	0.00	10.23	< .001
N strokes_C1	−0.02	0.00	−5.63	< .001
Ortho *N*_C1	0.00	0.00	−8.44	< .001
Ortho *N*_C2	0.00	0.00	−14.97	< .001
SR ortho *N*_C1	0.00	0.00	−3.44	.001
PR ortho *N*_C1	0.01	0.00	10.87	< .001
PR ortho *N*_C2	−0.01	0.00	−5.46	< .001
AoA	0.01	0.00	3.74	.028
Concreteness	0.06	0.01	8.64	< .001
Regularity_C1	0.04	0.01	6.57	< .001
Adjusted R^2 = .574				

Zipf_C1 CCLOWW frequency of the first character, *Zipf_C2* CCLOWW frequency of the second character, *N strokes* total number of strokes, *N strokes_C1* stroke counts 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; AoA = age of acquisition; Regularity_C1 = regularity of the first character. Only significant predictors are presented in the table

addition to SUBTLEX-CH frequencies in adults' visual word recognition. We examined this question by comparing the full model with the reduced models in which *zipf*_CCLOWW was first removed and then *zipf*_C1 and *zipf* _C2 were removed, using the naming and the lexical decision data. In word naming, CCLOWW word frequencies only contributed 0.75% additional variance, but the frequencies of the constituting characters explained an additional 13.10% variance in addition to SUBTLEX-CH word frequencies and other lexical variables. By contrast, in lexical decision, CCLOWW word frequencies explained 12.11% extra variance and frequencies of the constituting characters explained 1.28% extra variance in the latencies.

Discussion and conclusion

CCLOWW is the first grade-level simplified Chinese character and word database based on a large sample of reading materials for children. It contains three subcorpora intended for grade 2 and below, grades 3–4 and grades 5–6, providing grade-level frequency and contextual diversity statistics, as well as word lengths and syntactic categories data, for 6746 characters and 153,079 words, which were computed from a 34-million-character, 22-million-word corpus sampled from 2,152 documents. It is the largest children's written Chinese lexical corpus so far and the only one freely accessible online (https://www.learn2read.cn/database/).

Comparing character and word statistics of the subcorpora provides us some insights into the developmental trends of children's print exposure in Chinese. Compared with the lower grade subcorpus, novel characters take up a medium proportion of the G34 but a very small one of the G56 subcorpora. This suggests that children do not encounter many new characters beyond grades 3 and 4. By contrast, novel words account for considerable proportions of both the G34 and G56 subcorpora, indicating that children continue seeing many new words throughout the primary years. We have also found that across the subcorpora, words comprising two or more characters become increasingly frequent while the proportion of one-character words decreases. These findings are broadly in accord with the current literature suggesting that character knowledge develops prior to word knowledge, and that children read increasingly complex words as their reading experience grow (Su & Samuels, 2010). These developmental patterns provide support for designing teaching materials and pedagogies that takes into account what average children might see in print. The early years of primary education should devote more to teaching characters and short words, and the later years to teaching words with increasingly complex structures.

The frequency and contextual diversity measures correlated well among the subcorpora of CCLOWW. They also correlated well with a small child corpus (CJC, Huang et al., 2020) and with a large adult corpus (SUBTLEX-CH, Cai & Brysbaert, 2010), the latter of which has been considered a reliable reflection of adults' language experience in Chinese. Interestingly, correlations between CCLOWW and SUBTLEX-CH frequencies are slightly higher than that between CCLOWW and CJC, although CJC was built based on school textbooks and storybooks intended for children. This provides some support for Brysbaert and New's

(2009) proposal for a sizable corpus for computing reliable frequency statistics.

Through a series of validation analyses, we have confirmed that the frequency indices computed in CCLOWW are good estimates of children's print exposure in simplified Chinese. The character frequency and contextual diversity measures reliably predicted grade 2–3 children's character naming. Importantly, they outperformed the indices from the other children's database, CJC. Also, based on frequency measures from CCLOWW, we have demonstrated a significant frequency effect on grade 2–3 children's one- and two-word naming accuracy. These results have not only validated the measures in CCLOWW but have also shown a robust frequency effect in young children's visual word recognition. This is in line with findings of previous eye-tracking studies that children fixated longer on and were more likely to skip high-frequency than low-frequency Chinese words (Chen & Ko, 2011; Liu et al., 2021). Nevertheless, this result differs from that of Huang et al. (2020), which has found that the word frequency effect on children's lexical decision vanished once contextual diversity was taken into account. The theoretical reason for why CD might influence lexical processing is that the contextual experience of a word influences the likelihood of the word appearing in the future and thus its accessibility in the mental lexicon (Adelman et al., 2006; Jones et al., 2017). Yet, because it is highly correlated with frequency, there are concerns that it might simply be a better measure of frequency than frequency itself (Brysbaert & New, 2009). Indeed, word frequencies and CD are highly correlated in CCLOWW and their potential dissociative effects cannot be examined with the current data. Also, some recent studies have proposed that measures of CD should take into account semantic variability of a word in context, and such measures outperform frequency and document count in predicting naming and lexical decision latencies (Hsiao & Nation, 2018; Johns et al., 2015; Johns & Jones, 2022).Whether and how contextual diversity explains children's Chinese word processing requires further investigation in the future.

Analysis of the adults' word-naming and lexical decision RTs revealed that the character and word frequencies provided by CCLOWW could also explain adults' visual word processing. Among a number of orthographic (e.g., number of strokes), phonological (e.g., regularity) and semantic (e.g., concreteness) lexical and sublexical variables investigated in the analyses, CCLOWW frequencies of the words and of the constituting characters predicted adults' word-naming and lexical decision reaction times. Moreover, our exploratory analysis shows that CCLOWW word and character frequencies accounted for considerable extra variances in the adults' naming and lexical decision data, in addition to SUBTLEX-CH frequencies, possibly suggesting that print exposure in a reader's early years might have some long-term effects on their visual word processing. Interestingly, the contributions of CCLOWW word and character frequencies were reversed in naming and lexical decision. Word frequency did not contribute much variance in the naming RTs but it explained 12.11% extra variance in lexical decision. By contrast, frequencies of the constituting characters explained 13.10% extra variance in naming latencies but only 1.28% in lexical decision. This is compatible with findings of some previous works. For example, Morrison and Ellis (1995) found that the word frequency effect was diminished once other lexical variables were controlled in naming speed, but it remained significant in lexical decision RTs. In Chinese two-character word recognition, the word frequency effect was smaller in naming than in lexical decision (Gao et al., 2016). Our finding potentially indicates that although word naming and lexical decision both probe the process of visual word identification, naming might imply a less complete lexical access, at least in the recognition of two-character words. In a typical naming task, as soon as the participants pronounce the first character of the word, the reaction time is recorded. It is possible that the naming latency is affected more by the constituting character than the whole word. The word frequency effect in naming might thus be reduced. These proposals need to be investigated in future work.

One limitation of CCLOWW is that in constructing the corpus, although we ensured that the book assignment to the grade levels was supported by either recommendations of the Ministry of Education or by at least two other sources, it might not be optimal. We hope to provide a description of the features of characters and words children are exposed to at different developmental stages. This, of course, is contingent on a collection of reading materials arranged by grade or reading levels that children of that grade or level would indeed read. Nevertheless, leveled reading is currently not systematically available in mainland China, which resulted in this limitation. But it also motivated us to build a grade-level written word database, because we wanted to profile the reading materials children at different developmental stages are likely exposed to. Several indices calculated in CCLOWW, such as the average document length and numbers of unique character and word types, indeed show clear developmental changes, indicating that the documents in each subcorpora were close reflections of what children might read at that stage.

In sum, CCLOWW provides an important and reliable tool for research in children's reading and its development in simplified Chinese. By allowing for the selection of character and word stimuli adjusted to children's developmental stages, it could contribute to a wide range of research areas from building children's reading norms to developmental changes in cognitive and neural processes. Additionally, as CCLOWW provides characters and words statistics by grade level, it is a useful tool for educators, writers, and publishers, as it allows them to pick age-appropriate vocabulary when selecting and producing reading materials for children. In the future, we plan to extend CCLOWW to incorporate additional character and word variables that might also influence children's reading, such as grade-level phonological neighborhood, phonological frequency, concreteness and familiarity provided by children. We will also further validate the corpus by comparing the variance explained by the grade-level and cumulative frequency values in relation to naming and lexical decision latencies in children. The database will be updated as these new data are acquired.

Acknowledgements We would like to thank Ziyi Ding, Danni He and Fangzhou Liu for their help with proofreading the materials used for corpus construction.

Data and code availability The database is available at https://www.learn2read.cn/database/. Experimental data and code are available upon request.

Funding This work was supported by the National Natural Science Foundation of China (grant numbers 31970987) and by the Science and Technology Commission of Shanghai Municipality (19JC1410100).

Declarations

Conflicts of interest The authors declare no conflicts of interest.

Ethics approval The study was approved by the East China Normal University Committee on Human Research Protection. Written consent was obtained from the child participants' parents/guardians and the adult participants.

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.

Bai, X., Yan, G., Liversedge, S., & Zang, C. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1277–1287.

Balota, D. 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.

Bowey, J. A., & Muller, D. (2005). Phonological recoding and rapid orthographic learning in third-graders' silent reading: A critical test of the self-teaching hypothesis. *Journal of Experimental Child Psychology, 92*(3), 203–219.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Mandera, 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

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE, 5*(6). https://doi.org/10.1371/journal.pone.0010729

Chen, M., & Ko, H. (2011). Exploring the eye-movement patterns as Chinese children read texts: a developmental perspective. *Journal of Research in Reading, 34*(2), 232–246. https://doi.org/10.1111/j.1467-9817.2010.01441.x

Chen, H. C., Song, H., Lau, W. Y., Wong, K. F. E., & Tang, S. L. (2003). Developmental characteristics of eye movements in reading Chinese. In C. McBride-Chang & H.-C. Chen (Eds.), *Reading Development in Chinese Children* (pp. 157–169). Praeger.

Chen, B., Dent, K., You, W., & Wu, G. (2009). Age of acquisition affects early orthographic processing during Chinese character recognition. *Acta Psychologica, 130*(3), 196–203. https://doi.org/10.1016/j.actpsy.2008.12.004

Cunningham, A., & Stanovich, K. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6), 934–945.

Ehri, L. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*(2), 167–188.

Gao, X. Y., Li, M. F., Chou, T. L., & Wu, J. T. (2016). Comparing the frequency effect between the lexical decision and naming tasks in Chinese. *Journal of Visualized Experiments*, (110), e53815. https://doi.org/10.3791/53815

Geng, Z., Yan, H., Qiu, X., & Huang, X. (2020). fastHan: A BERT-based Multi-Task Toolkit for Chinese NLP. *ArXiv Preprint*, arXiv:2009.08633. Retrieved from https://arxiv.org/abs/2009.08633

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.

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

Johns, B., Dye, M., & Jones, M. (2015). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review, 23*(4), 1214–1220.

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

Keuleers, E., Brysbaert, 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

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

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39*(2), 192–198.

Liu, N., Wang, X., Yan, G., & Paterson, K. B. (2021). Eye Movements of Developing Chinese Readers: Effects of Word Frequency and Predictability. *Scientific Studies of Reading, 25*(3), 234–250. https://doi.org/10.1080/10888438.2020.1759074

Martínez Martín, J. A., & García Pérez, M. E. (2008). ONESC: A database of orthographic neighbors for Spanish read by children. *Behavior Research Methods, 40*(1), 191–197. https://doi.org/10.3758/BRM.40.1.191

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2003). *Children's printed word database*. Available from: http://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.

McFadden, D. (1977). Quantitative Methods for Analysing Travel Behaviour of Individuals. In D. Hensher & P. Stopher (Eds.), *Bahavioural Travel Modelling* (pp. 279–318). Routledge.

Ministry of Education, R. O. C. (2013). *Table of General Standard Chinese Characters*. Retrieved October 01, 2021, from http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm

Ministry of Education, R. O. C. (2019). *2019 Book recommendation for primary and middle school libraries*. Retrieved from http://www.moe.gov.cn/srcsite/A06/s3321/201911/W020191112396369462367.pdf

Ministry of Education, R. O. C. (2020). *2020 Extracurricular reading recommendation for primary and middle school students*. Retrieved from http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202004/W020200422556593462993.pdf

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*(2), 267–296. https://doi.org/10.1037/a0021890

Morrison, C., & Ellis, A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 21*(1), 116–133.

National Press and Publication Administration. (2020). *2020 Annual Report of Digital Reading in China. Beijing*. Retrieved from http://www.nppa.gov.cn/nppa/contents/280/75940.shtml

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

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

Share, D. L. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology, 87*(4), 267–298.

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.

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*(1), 240–253. https://doi.org/10.3758/s13428-013-0350-1

Su, Y. F., & Samuels, J. J. (2010). Developmental changes in character-complexity and word-length effects when reading Chinese script. *Reading and Writing, 23*(9), 1085–1108. https://doi.org/10.1007/S11145-009-9197-3

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

Sze, W. P., Yap, M. J., & Rickard Liow, S. J. (2015). The role of lexical variables in the visual recognition of Chinese characters: A megastudy analysis. *The Quarterly Journal of Experimental Psychology, 68*(8), 1541–1570.

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.

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

Tse, C.-S., & Yap, M. J. (2017). The role of lexical variables in the visual recognition of two-character Chinese compound words: A megastudy analysis. *Quarterly Journal of Experimental Psychology, 71*(9), 2022–2038. https://doi.org/10.1177/1747021817738965

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

Wang, H., Chang, R. B., & Li, Y. S. (1986). *Modern Chinese Frequency Dictionary*. Beijing Language Institute.

Xing, H., Shu, H., & Li, P. (2004). The acquisition of Chinese characters: Corpus analyses and connectionist simulations. *Journal of Cognitive Science, 5*(1), 1–49.

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

Xu, X., Li, J., & Guo, S. (2021). Age of acquisition ratings for 19,716 simplified Chinese words. *Behavior Research Methods, 53*, 558–573. https://doi.org/10.3758/s13428-020-01455-8

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

Yan, G., Tian, H., Bai, X., & Rayner, K. (2006). The effect of word and character frequency on the eye movements of Chinese readers. *British Journal of Psychology, 97*(2), 259–268.

Zhang, B., & Peng, D. (1992). Decomposed storage in the Chinese lexicon. *Advances in Psychology, 90*, 131–149.

Zipf, G. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

Open practices statement The database is available at https://www.learn2read.cn/database/. Experimental data and code are available upon request. None of the experiments were preregistered.