


ORIGINAL ARTICLE

The processing of multiword expressions in children and adults: An eye-tracking study of Chinese

Shang Jiang¹, Xin Jiang² and Anna Siyanova-Chanturia^{1,3,*} 

¹Te Herenga Waka - Victoria University of Wellington, ²Beijing Language and Culture University and

³Ocean University of China

*Corresponding author. E-mail: anna.siyanova@vuw.ac.nz

(Received 3 May 2019; revised 1 May 2020; accepted 15 May 2020; first published online 24 August 2020)

Abstract

The processing advantage for multiword expressions over novel language has long been attested in the literature. However, the evidence pertains almost exclusively to multiword expression processing in adults. Whether or not other populations are sensitive to phrase frequency effects is largely unknown. Here, we sought to address this gap by recording the eye movements of third and fourth graders, as well as adults (first-language Mandarin) as they read phrases varying in frequency embedded in sentence context. We were interested in how phrase frequency, operationalized as phrase type (collocation vs. control) or (continuous) phrase frequency, and age might influence participants' reading. Adults read collocations and higher frequency phrases consistently faster than control and lower frequency phrases, respectively. Critically, fourth, but not third, graders read collocations and higher frequency phrases faster than control and lower frequency sequences, respectively, although this effect was largely confined to a late measure. Our results reaffirm phrase frequency effects in adults and point to emerging phrase frequency effects in primary school children. The use of eye tracking has further allowed us to tap into early versus late stages of phrasal processing, to explore different areas of interest, and to probe possible differences between phrase frequency conceptualized as a dichotomy versus a continuum.

Keywords: children; Chinese; eye movements; language processing; multiword expressions

The online processing of multiword expressions (MWEs), and phrase frequency effects in particular, have figured prominently in psycholinguistic literature in the past decade (for a review, see Siyanova-Chanturia & van Lancker Sidtis, 2019). A plethora of recent studies have firmly attested to a key role of phrase frequency in language processing (on a par with lexical frequency, e.g., Balota & Chumbley, 1984). Despite the different approaches to data collection and analysis, the evidence that has so far accumulated suggests that proficient (adult) language users are highly sensitive to the distributional properties of multiword information, both at the level of

comprehension (e.g., Arnon & Snider, 2010; Siyanova-Chanturia, Conklin, & Schmitt, 2011; Siyanova-Chanturia, Conklin, & van Heuven, 2011; Tremblay & Baayen, 2010; Tremblay, Derwing, Libben, & Westbury, 2011) and production (e.g., Arnon & Cohen Priva, 2013, 2014; Bell *et al.*, 2003; Bybee & Scheibman, 1999; Janssen & Barber, 2012; Siyanova-Chanturia & Janssen, 2018; Tremblay & Tucker, 2011). Overall, MWEs have been found to be processed differently from novel language. While behavioral studies (above) generally show faster processing (e.g., reading and articulation), electrophysiological studies point to easier semantic integration and template matching mechanisms for MWEs relative to novel phrases (e.g., Molinaro & Carreiras, 2010; Siyanova-Chanturia, Conklin, Caffarra, Kaan, & van Heuven, 2017; Vespignani, Canal, Molinaro, Fonda, & Cacciari, 2010).

While processing studies with adult speakers abound, very little is known about MWE processing in children. This is rather surprising given the important role that has long been attributed to sequences above the word level in first language (L1) learning and use. Studies into L1 acquisition have attested to the interplay between holistic and analytic language processes (e.g., Locke, 1997; Peters, 1977, 1983; Tomasello, 2003). As Bolinger (1975) noted, L1 learning is initially holistic, only later becoming more analytical. Researchers have long documented that young children produce unanalyzed chunks (along with single words), such as *Lemme see*, *I wanna do it*, and *Gimme that* (e.g., Clark, 1974; Cruttenden, 1981; Lieven, Pine, & Barnes, 1992; Nelson, 1973; Peters 1983). What this implies is that children learn linguistic structures of different shapes and sizes and of various degrees of abstraction (e.g., Tomasello, 2003).

Although researchers largely agree that young children are capable of memorizing and using relatively complex strings of words before they are capable of analyzing their internal structure, the role attributed to chunks in the early (naturalistic) studies has not always been viewed positively. For example, Bates, Bretherton, and Snyder (1988) viewed such strings as linguistic “dead-ends.” Similarly, Brown and Hanlon (1970) argued that because unanalyzed chunks resist segmentation, they are unlikely to contribute to the child’s linguistic development. As a result, children whose early vocabularies are characterized by memorized chunks were viewed as slow learners unable to analyze and segment adult speech (e.g., Bates *et al.*, 1988; Bretherton, McNew, Snyder, & Bates, 1983). In contrast, other researchers have proposed that chunks in children’s speech play a crucial role in their early linguistic development (e.g., Clark, 1974; Peters, 1977, 1983; Pine & Lieven, 1993; Tomasello, 2003; Tomasello & Brooks, 1999). For example, Lieven *et al.* (1992) found that the use of chunks correlated positively with general productivity (also see Clark, 1974; Cruttenden, 1981). Similarly, Peters (1977) suggested that children are capable of breaking down multiword strings into constituent components, a process thought to contribute directly to the development of adultlike morphosyntax.

Research that has since followed has overwhelmingly supported the idea that MWEs play a pivotal role in the process of L1 learning. As Tomasello (2003) put it, the existence of chunks, or “holophrases,” is “of tremendous theoretical importance for theories of linguistic competence and performance” (p. 306). What the above evidence further alludes to is that children remember utterances they are exposed to frequently and are able to store multiword information along with single words (e.g., Goldberg, 2006; Tomasello, 2003).

Despite the important role attributed to sequences above the word level in L1 learning and use, the majority of studies to date have been based on naturalistic observations. Experimental evidence with L1 children is extremely scarce. To the best of our knowledge, only one published study has looked at the role of phrase frequency in online language processing in children. Using a repetition (production) task, Bannard and Matthews (2008) set out to test whether young children store and reuse multiword sequences encountered in their input. These authors examined the accuracy and speed with which 2- and 3-year-old children produced four-word sequences. To this aim, 13 pairs varying in frequency (e.g., *a lot of noise*, log frequency = 4.66 vs. *a lot of juice*, log frequency = 0.69) were created using the Max Planck Child Language Corpus. The longitudinal corpus contains the speech produced by a male child between the ages of 2 and 5, as well as his mother's speech addressed to him. The target stimuli were created based on the latter, given the authors' interest in the kind of language the child was exposed to. Four-word sequences were chosen because these allowed a wide enough frequency range, and because these (rather than shorter) sequences were sufficiently long to observe the variance in the children's performance. The phrases within each pair were controlled for the frequency of the final word (e.g., *noise* vs. *juice*), the frequency of the final bigram (e.g., *of noise* vs. *of juice*), and the length of the final word in syllables. Bannard and Matthews (2008) found that children as young as 3 were sensitive to phrase frequency distributions, with more frequent phrases being articulated more quickly than less frequent phrases. This study further showed that children as young as 2 were more likely to repeat a word sequence correctly if it was higher rather than lower frequency. Thus, the frequency of the target phrasal configurations (as attested in the reference corpus) affected the speed as well as accuracy with which 2- and 3-year-olds produced them, suggesting that young children possess "experience-derived knowledge of specific four-word sequences" (p. 246). The knowledge of multiword sequences, it was argued, was in addition to the children's knowledge of the individual constituents that make up the sequences. The authors concluded that the children possessed "complementary representations at different levels of granularity" (Bannard & Matthews, 2008, p. 246).

Bannard and Matthews (2008) took their results to support the usage-based and exemplar-based approaches to language acquisition, processing, and use, which reject the lexicon-grammar dichotomy and argue for a central role of frequency of exposure in L1 learning (e.g., Abbot-Smith & Tomasello, 2006; Bod, 2006; Bybee, 1998; Goldberg, 2006; Tomasello, 2003). A key assumption behind these models is that the basic unit of language acquisition is a construction (e.g., Goldberg, 2006; Tomasello, 2003), and that children are able to extract recurrent sequences—varying in length, complexity, and level of abstractness—from the available input, store them, and subsequently use them in their output (e.g., Lieven, Behrens, Speares, & Tomasello, 2003).

The present study

The findings presented in Bannard and Matthews (2008) are important as they attest to children's sensitivity to the distributional properties of multiword information. However, the evidence is limited to the *repetition* of MWEs, that is, elicited

(controlled) production out of meaningful sentential context (which, admittedly, was the only task possible given the age of the participants). Further, the study's focus was on very young L1 learners; how older children deal with multiword information online has not been investigated. The primary aim of the present study was, thus, to forge a better understanding of phrase frequency effects in L1 children's online processing by focusing on language comprehension and employing a naturalistic reading task. For the purpose of this study, and in line with earlier research (e.g., Bannard & Matthews, 2008), we adopted the frequency-based approach to the treatment of MWEs. That is, we relied primarily on the frequency of occurrence of MWEs in a representative corpus when identifying and selecting the target items.

In addition, we decided to use eye movements due to their vast applications in reading research (e.g., Rayner, 1998, 2009). First, this method allows for separate analyses to be performed on early, middle, and late stages of reading (for an overview of these measures in the context of vocabulary and MWE research, see Pellicer-Sánchez & Siyanova-Chanturia, 2018; Roberts & Siyanova-Chanturia, 2013; Siyanova-Chanturia, 2013). While early measures (e.g., first fixation duration) are sensitive to early processes, such as lexical access and early integration of information, late measures (e.g., dwell time and fixation count) are associated with later processing mechanisms, such as information (re)analysis and discourse integration (e.g., Paterson, Liversedge, & Underwood, 1999; Rayner, Sereno, Morris, Schmauder, & Clifton, 1989). A combination of both early and late measures provides the researcher with an extremely rich picture of one's reading behavior. Second, during the course of the experiment, readers do not need to perform a secondary task, such as deciding whether a string of letters is a real word or a nonword (although it is still advisable to provide comprehension questions, see Method section). Third, unlike other experimental paradigms commonly used in reading research (e.g., self-paced reading, rapid serial visual presentation, etc.), the text can be presented over an entire screen. As a result, the eye movement methodology is as close to natural reading as possible in a laboratory setting (e.g., Duyck, van Assche, Drieghe, & Hartsuiker, 2007), an important consideration given the age of our participants.

Further, in order to extend available evidence to other age groups and L1 backgrounds, we targeted Chinese (L1 Mandarin) primary school pupils, in particular, third and fourth graders (8- and 9-year-olds, respectively), along with adults. Although the 3-year-old children in Bannard and Matthews (2008) were better at repeating sequences than their 2-year-old counterparts, these authors found no significant interaction between children's age and phrase frequency, suggesting "continuity in frequency effects across development" (p. 247). We wanted to probe this assertion further.

We decided to focus on Chinese, a language markedly different from English. First, unlike alphabetic languages (such as English), Chinese is a logographic language with characters varying in their complexity and number of strokes, a factor known to affect fixation durations in eye movement studies (e.g., Ma & Li, 2015). It has been proposed that reading (as well as writing) Chinese characters may require a unique set of skills that are different from alphabetic languages (e.g., Chung & McBride-Chang, 2011), due to the "visual complexity" of characters and their large number (Liu, Chen, & Chung, 2015, p. 307). Second, many Chinese characters are homophones. According to Chao (1976), Li, Anderson, Nagy, and Zhang (2002)

and Li, Wang, Tong, and McBride (2017), among the commonly taught 7,000 Chinese characters, there are only 1,277 syllables with different tones, meaning that “about 5 different characters share the same pronunciation” (Li et al., 2017, p. 142). According to these authors, this feature makes Chinese scripts rather complex in terms of literacy demands and learning to read. Third, one of the major ways in which Chinese differs from English is word segmentation. According to Packard (2000), a Chinese character can stand alone as a word, can be combined with another character to form a two-character word (which is very common), or can be combined with two or more characters to form a longer multicharacter word (which is less common). This variability may further add to the ambiguity around word boundaries and word segmentation during reading (Liu, Li, Lin, & Li, 2013). In addition, unlike English and other European languages, Chinese does not mark word boundaries with spaces. Although “readers of Chinese are able to perform word-segmentation analysis during parafoveal previewing,” they may do it “less efficiently” than, for example, English readers (Packard, 2016, p. 316).

With respect to MWE processing in Chinese, the evidence is rather limited, with fewer than a handful of published studies to date. Using a grammaticality judgment task, Kong, Zhang, and Zhang (2016) found that higher frequency discontinuous correlative conjunctions (e.g., 因为...所以... “because ... therefore ...”; 虽然...也... “although ... yet...”) were read faster than lower frequency phrases. More recently, Yi, Lu, and Ma (2017) recorded eye movements of L1 and second language (L2) speakers of Chinese as they read two-word adverbial sequences in sentential context. These authors found that both groups of readers were sensitive to phrase frequency and contingency (operationalized as mutual information, a measure that shows how likely the two words co-occur together) of target MWEs. Contrary to earlier research (e.g., Siyanova-Chanturia, Conklin, & van Heuven, 2011), L2 speakers exhibited greater sensitivity to phrase frequency compared to L1 speakers. Finally, Yu et al. (2016) investigated parafoveal preview effects in Chinese by using idioms versus control phrases. Although this was not their primary line of enquiry, Yu et al. (2016) reported faster reading for idioms than control phrases. It is interesting that there was no indication that idioms benefited from greater parafoveal processing than controls. Overall, albeit scarce, the evidence points to a processing advantage for Chinese MWEs over novel strings of language, akin to what has been reported in the literature on English and other languages.

The main focus of the present study was on phrase frequency effects (or lack thereof) in primary school children. Phrase frequency, however, can be and has been operationalized in a number of different ways. Target phrases have been assigned to various frequency bins (e.g., Sosa & MacFarlane, 2002), have been looked at as a dichotomy (i.e., high frequency vs. low frequency; e.g., Jiang & Nekrasova, 2007), or have been treated as a continuum (e.g., Arnon & Snider, 2010). Some studies have included a continuous as well as a binary (high vs. low) measure of frequency, reporting a distinct pattern of results for the two (e.g., Siyanova-Chanturia, Conklin, & van Heuven, 2011; Siyanova-Chanturia & Janssen, 2018). It has also been proposed that a continuous measure of frequency may be a better predictor of response times compared to employing frequency bins (e.g., Arnon & Snider, 2010). With this in mind and to further probe this assertion, we decided to

include both a continuous and a dichotomous measure of phrase frequency, hoping to obtain a richer picture of phrase frequency effects in MWE processing in children.

Based on the literature reviewed above and the gaps outlined, we sought to answer the following questions:

1. Are Chinese adults sensitive to phrase frequency during reading?
2. Are Chinese third and fourth graders sensitive to phrase frequency during reading?
3. Do reading patterns reveal any processing differences between the effects of phrase frequency conceptualized as a dichotomy versus a continuum?

Method

Participants

Thirty-five normally developing pupils (17 males and 18 females) from two primary schools in Beijing participated in the study. Nineteen of them (9 males and 10 females) were in Grade 3 (age range: 8 years and 3 months to 9 years, mean age = 8 years and 7 months), and 16 of them (8 males and 8 females) were in Grade 4 (age range: 9 years and 4 months to 10 years, mean age = 9 years and 6 months). The Grade 3 primary school students were selected because at this age, children are considered to have mastered enough characters to read reasonably well (Wu *et al.*, 2009). The fourth graders were selected because at this age, children's comprehension processes start to approximate those of adult readers (Rayner, Pollatsek, Ashby, & Clifton, 2012). In addition, 26 adult participants (11 males and 15 females) were recruited from the student population of the Beijing Language and Culture University (BLCU; age range: 18 years to 30 years old, mean age = 21 years and 5 months). All participants were native speakers of Mandarin and had normal or corrected-to-normal vision. They received a small gift for their participation. Written consent was obtained from the adult participants, and from the children's parents and their teachers. The study was carried out in line with the ethical procedures of the BLCU.

Materials

The BCC corpus of Chinese¹ (BLCU Corpus Center, 2016) was used to select experimental items as well as to extract their phrase and lexical frequencies. Forty pairs (80 items in total) of verb–noun combinations varying in phrase frequency were extracted from the corpus. Forty items were high-frequency phrases (e.g., 参加会议 “attend the meeting,” frequency = 2141), and forty were low-frequency phrases (e.g., 参加游戏 “attend the game,” frequency = 62). Thus, each item was assigned to either a high-frequency (collocation) condition or a low-frequency (control) condition. All phrases were transparent, literal and fully compositional. Nouns and verbs in both sequence types consisted of two Chinese characters each, while the entire phrasal configuration was always four characters in length. The verb was always identical in the two conditions (e.g., 参加 “attend”).

The nouns in the collocation and control conditions were matched for frequency (e.g., 会议 “meeting” in collocation and 游戏 “game” in control; $Mean_{noun}$ in collocation = 96369.73, $Mean_{noun}$ in control = 108061.83, $t = 0.79$, $p > .10$; $Median_{noun}$ in collocation = 83610, $Median_{noun}$ in control = 72933; $Mean \log frequency_{noun}$ in collocation = 11.17, $Mean \log frequency_{noun}$ in control = 11.20). The stroke number and structure of the two characters in the noun were also matched (Character 1 stroke number: $t_{paired} = -0.66$, $p > .10$; Character 2 stroke number: $t_{paired} = -0.76$, $p > .10$; Character 1 structure: $\chi^2 = 5.63$, $p > .10$; Character 2 structure: $\chi^2 = 16.28$, $p = .09$). The two types of phrases were, thus, closely matched for verb and noun frequency, phrase length, character structure, and stroke number. However, the collocations and their controls differed significantly in phrase frequency ($Mean_{collocation} = 2071.30$, $Mean_{control} = 16.08$, $t = -3.77$, $p < .001$; $Median_{collocation} = 1005.50$, $Median_{control} = 4$; $Mean \log frequency_{collocation} = 6.97$, $Mean \log frequency_{control} = 1.81$). The constituent words within the target phrases were all high-frequency words (frequency ≥ 1639) according to the BCC corpus, an important consideration given the age of the participants.² The two words in the collocation and control conditions were matched for (forward) association strength (see Siyanova-Chanturia, Conklin, & van Heuven, 2011; and Siyanova-Chanturia et al., 2017, for a similar procedure). To this aim, 10 adult native speakers of Chinese, who did not participate in the eye-tracking experiment, took part in the norming procedure. Participants were asked to provide the first word that came to mind after seeing Word 1 of a phrase (i.e., verb, e.g., 参加 “attend”). The proportion of the nouns (out of the 10 responses obtained) identical to the noun in the target item (e.g., 会议 “meeting” in collocation, or 游戏 “game” in control) was used as the association strength score for the phrase. *T* test showed no significant differences in the strength of association between Word 1 and Word 2 in collocations versus controls ($t = -0.95$, $p > .10$). The characteristics of the experimental items are summarized in Table 1.

Collocations and their respective controls were embedded in an identical sentence context. The length of the sentences ranged from 12 characters to 21 characters (mean = 15.35 characters), with the target items appearing roughly in the middle of the sentence, at least three Chinese characters from the end of the sentence. An example of a sentence in the collocation and control conditions is provided below (target phrase underlined):

Collocation:	这次来 <u>参加会议</u> 的学生有10人。 “Ten students <u>attended the meeting.</u> ”
Control:	这次来 <u>参加游戏</u> 的学生有10人。 “Ten students <u>attended the game.</u> ”
Comprehension question:	没有学生来参加会议 / 游戏吗? “No students attended the meeting/game?”

Procedure

Before the reading experiment, the background information regarding each participant (e.g., gender and age) was recorded. All child participants were tested on their

Table 1. Summary of phrase frequency, word frequency, and association strength for the target items

Phrase type	Example	Min phrase frequency	Max phrase frequency	Mean phrase frequency	Median phrase frequency	Mean log phrase frequency	Mean frequency of word1 (verb)	Mean frequency of word 2 (noun)	Median frequency of word 2 (noun)	Mean log frequency of word 2 (noun)	Mean association strength
Collocation	参加会议 (attend the meeting)	114	20800	2071.30	1005.50	6.97	35271.08	96369.73	83610	11.17	0.15
Control	参加游戏 (attend the game)	1	126	16.08	4.00	1.81	35271.08	108061.83	72933	11.20	0.12

Chinese character reading.³ Results showed that all child participants were able to read/recognize the top 2,000 most frequently used characters, and that Grade 4 (scores ranged from 81 to 100, mean = 94.38) significantly outperformed Grade 3 (scores ranged from 78 to 97, mean = 89.05) in this test ($t = -3.23$, $p < .01$, effect size = 1.10).

The reading experiment was run individually in the Cognition & Neuroscience Lab at the BLCU, using the Eyelink 1000 plus (SR Research Ltd., 2016). A 9-point grid calibration procedure was conducted before the experiment. Participants first completed a practice session, which included six trials. Each trial started with a fixation point that appeared at the beginning of the upcoming sentence. After participants saw the fixation point, a sentence appeared across one line in the middle of the screen. Participants were asked to read as quickly as possible for comprehension. Each trial was followed by a comprehension question (e.g., see Tremblay et al., 2011). An example of a comprehension question is presented above.

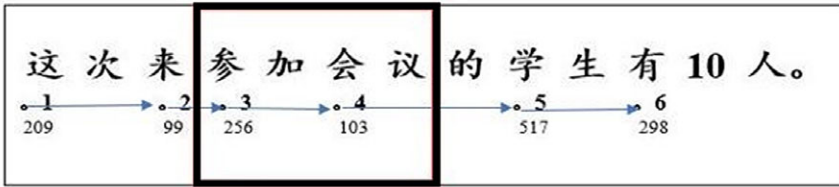
Two counter-balanced lists were used with the items randomized (each list included 20 collocation sentences and 20 control sentences). Only one condition of the phrase (either collocation or control) was seen by each participant. After the experiment, the participants (children and adults) were asked to rate the naturalness of the collocation and control sentences on a 10-point Likert scale (请用数字1–10对下列句子的自然度进行评分。“Please rate the naturalness of the following sentences on a scale of 1–10, where 1 is unnatural and 10 is natural.” 例如: 杯子懂得珍惜时间 [1, “Cups know how to cherish time”]; 月亮在天上保护着你 [5, “The moon in the sky is protecting you”]; 妈妈正在房间里打扫卫生 [10, “The mother is cleaning the room”]). The experiment took about 30 min from start to finish.

Data analysis and results

Based on the results of the naturalness rating task, we decided to exclude five pairs of sentences from the final data analyses. No significant differences in naturalness judgments were found between the remaining 35 pairs of sentences (total: $Mean_{\text{collocation-sentences}} = 8.50$, $Mean_{\text{control-sentences}} = 8.30$, $t = -1.27$, $p > .10$; Grade 3: $Mean_{\text{collocation-sentences}} = 8.63$, $Mean_{\text{control-sentences}} = 8.46$, $t = -1.07$, $p > .10$; Grade 4: $Mean_{\text{collocation-sentences}} = 8.46$, $Mean_{\text{control-sentences}} = 8.24$, $t = -1.36$, $p > .10$; adults: $Mean_{\text{collocation-sentences}} = 8.43$, $Mean_{\text{control-sentences}} = 8.22$, $t = -1.24$, $p > .10$).

Single fixation durations shorter than 80 ms or longer than 1000 ms were discarded (e.g., Gu & Li, 2015; Wei, Li, & Pollastek, 2013; Zhou, Ma, Li, & Taft, 2017). The data loss accounted for 6.51% of the total data (3.21% for collocation sentences and 3.30% for control sentences). Given that including skipped items in the calculation of means or in any further analyses for a duration measure is problematic for a single-word analysis (e.g., Conklin, Pellicer-Sánchez, & Carrol, 2018; also see Vilkaitė, 2016), we discarded trials with a first fixation of 0 ms in the phrase-final word analysis. This exclusion accounted for 7.73% of the entire data (3.79% for the phrase-final word in collocation sentences, and 3.93% for the phrase-final word in

Collocation-embedded sentence:



Control-embedded sentence:

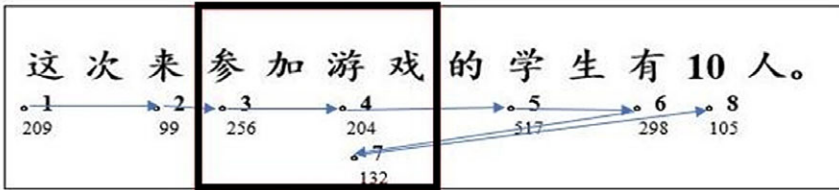


Figure 1. An example of a hypothetical eye-movement record for a collocation-embedded sentence and its control. The black box represents the area of interest (English translation: collocation = “attend the meeting,” control = “attend the game”). First fixation duration = 3, first pass reading time = 3 + 4, dwell time = 3 + 4 + 7 (refers to the sum of fixation durations), fixation count = 3 + 4 + 7 (refers to the sum of all fixations).

control sentences). The participants had no difficulty answering comprehension questions (Grade 3, 93.98% correct; Grade 4, 95.71% correct; adults, 97.58% correct).

A mixed design was employed, treating *phrase type* as a within-group variable, *age group* as a between-group variable, and the *eye-tracking measures* as dependent variables. In line with Carrol and Conklin (2019), Sonbul (2014), and Yi et al. (2017), four measures were examined: first fixation duration (early), first pass reading time (middle), dwell time, and fixation count (late, see Figure 1). The former three measures were *duration* measures, which reflected reading times, while the latter measure was a *count* measure, which provided the number of fixations.

In line with previous literature (e.g., Carrol & Conklin, 2014, 2015, 2019; Carrol, Conklin, & Gyllstad, 2016; Conklin & Pellicer-Sánchez, 2016; Underwood, Schmitt, & Galpin, 2004; Vilkaitė, 2016; Vilkaitė & Schmitt, 2017), two interest areas were selected for data analysis: the phrase-final word (e.g., 会议 “meeting”) and the whole phrase (e.g., 参加会议 “attend the meeting”). It was important to include the final word as an interest area, as well as the whole phrase, because the final word in MWEs “may be predictable and therefore elicit shorter fixations and more skipping” (Conklin, Pellicer-Sánchez, & Carrol, 2018, p. 102). Carrol and Conklin (2014) further recommend including both of these two interest areas, so as to capture “both the macro and micro features of formulaic units” (p. 4).

Descriptive statistics for the two experimental conditions across the four eye-tracking measures in the two interest areas can be found in Table 2. The data were analyzed using linear mixed-effects models in R (version 3.5.1, 2018), package *lme4* (version 1.1–21, 2019). The *p* value for each predictor (variable) was estimated by

Table 2. Mean fixation durations (in milliseconds) and mean fixation counts for the collocation (COL)-embedded sentences and their controls (CON) in analyses of the phrase-final word and the whole phrase, with standard deviation (SD) in parentheses

	Grade 3			Grade 4			Adults		
	COL sentence	CON sentence	<i>p</i> value	COL sentence	CON sentence	<i>p</i> value	COL sentence	CON sentence	<i>p</i> value
<i>Final word</i>									
FFD	409.76 (182.28)	406.21 (167.70)	.60	339.41 (155.92)	366.60 (178.08)	.03	283.71 (139.10)	319.11 (152.85)	<.001
FPRT	450.23 (228.88)	459.39 (260.83)	.32	354.18 (175.68)	387.67 (198.43)	.02	291.57 (154.72)	330.18 (171.78)	<.001
DT	731.30 (515.36)	742.12 (559.64)	.40	546.40 (379.39)	602.63 (463.71)	.07	343.75 (219.43)	414.75 (270.33)	<.001
FC	2.85 (1.75)	2.88 (1.77)	.43	2.36 (1.54)	2.48 (1.60)	.20	1.48 (0.76)	1.72 (0.99)	<.001
<i>Whole phrase</i>									
FFD	237.94 (86.01)	240.50 (102.72)	.37	229.95 (98.64)	239.08 (82.05)	.12	227.59 (76.95)	234.30 (87.14)	.11
FPRT	605.18 (482.88)	630.01 (573.01)	.28	469.78 (379.95)	506.69 (389.87)	.13	341.78 (220.05)	408.24 (275.40)	<.001
DT	1557.75 (1039.68)	1642.39 (1060.19)	.15	1197.59 (834.47)	1354.78 (859.33)	.02	612.17 (341.95)	834.05 (391.29)	<.001
FC	6.24 (3.12)	6.32 (3.51)	.38	5.85 (2.82)	6.05 (2.86)	.20	2.75 (1.45)	3.20 (1.72)	<.001

Notes: The mean fixation durations and fixation counts are the original values (not log-transformed). The estimated value of each predictor was rounded and kept to two decimal places. FFD, first fixation duration. FPRT, first pass reading time. DT, dwell time. FC, fixation count.

using the *lmerTest* package (version 3.1–0, 2019). The mixed-effects models for the three duration measures (i.e., first fixation duration, first pass reading time, and dwell time) were constructed using the *lmer* function (in the *lme4* package). As fixation count is discrete data, it needs to be analyzed differently from duration measures (e.g., Conklin *et al.*, 2018). Thus, we constructed a generalized linear model (*glmer* function in the *lme4* package) for this measure, specifying the “Poisson” distribution when fitting the model (e.g., Conklin *et al.*, 2018, p. 201; also see Vilkaitė, 2016). All the continuous reading measures and frequency measures were log-transformed. Kappa value was used to check for collinearity among predictors (variables) for each model.

To address the issue of collinearity between phrase type and phrase frequency ($\kappa = 107.43$, $corr = 0.86$), we built separate models for the two variables for each of the eye-tracking measures in the phrase-final word and whole phrase analysis. Thus, the two variables were not included in the same model.

The categorical variable *phrase type* included two conditions, collocation and control, which were coded as 1 (*collocation*) and –1 (*control*). *Age group* was also a categorical variable, coded as 3 for *Grade 3*, 4 for *Grade 4*, and 5 for adults.

Sixteen separate models were built: 2 Models (one with *phrase type* and one with *phrase frequency*) \times 4 Eye-Tracking Measures \times 2 Interest Areas. To build the original (maximal) models, *phrase type* (or *phrase frequency*), *age group* as well as the *interaction* between them were first added to the model, serving as primary predictors (fixed effects) for the reading times (or fixation count), regardless of the significance values. A maximal structure of random effects was then added to the model (Barr, Levy, Scheepers, & Tily, 2013), with random intercepts for participants and items, as well as random slopes for the main predictors (variables) involved.

In the model fitting procedure, we followed Bates, Kliegl, Vasishth, and Baayen (2015), removing the predictors stepwise from the full (maximal) model to a final converged one, in which the removal of a given predictor was no longer justified (e.g., Siyanova-Chanturia & Janssen, 2018). The justification for the removal was determined by model comparisons by using the chi-square tests in the *anova* function (in the *lmerTest* package). The *ranova* function (in the *lmerTest* package) was also checked for the random effect removal and model reduction.

Further comparisons/contrasts between the levels within a categorical variable (e.g., *age group* or *phrase type*) and interactions with other variables including calculating effect sizes for each comparison/contrast were performed using *emmeans* package (version 1.4.2, 2019). Further analyses of a continuous variable (e.g., *phrase frequency*) and interactions with other variables were conducted using the *aov* function in the *stats* package (version 3.6.1, 2019). The effect size for *aov* was calculated using the *etaSquared* function in the *lsr* package (version 0.5, 2019). The interactions were plotted by applying the *ggplot2* package (version 3.2.1, 2019).

Analysis of the phrase-final word

Fixed effects and random effects of selected models for the four eye-tracking measures, first fixation duration (FFD), first pass reading time (FPRT), dwell time (DT), and fixation count (FC), are presented in Table 3 (models with phrase type) and Table 4 (models with phrase frequency). *Phrase type* was not a significant predictor

Table 3. Summary of selected models (with phrase type) in the phrase-final word analysis

Fixed effects	First fixation duration					First pass reading time					Dwell time					Fixation count			
	Estimate	SE	df	t	pr	Estimate	SE	df	t	pr	Estimate	SE	df	t	pr	Estimate	SE	z	pr
Intercept	6.41	0.12	58.97	55.26	<.001	6.57	0.13	59.37	49.49	<.001	7.32	0.18	64.56	39.82	<.001	1.96	0.15	13.10	<.001
Age group	-0.17	0.03	58.05	-6.01	<.001	-0.19	0.03	58.35	-6.16	<.001	-0.31	0.04	60.09	-7.30	<.001	-0.30	0.03	-8.77	<.001
Phrase type	0.06	0.05	1644.73	1.18	.24	0.06	0.05	1590.88	1.11	.27	0.11	0.08	65.74	1.43	.16	0.12	0.07	1.58	.11
Age group × Phrase type	-0.02	0.01	1796.69	-1.95	.05	-0.02	0.01	1806.47	-1.94	.05	-0.04	0.02	65.28	-2.42	.02	-0.04	0.02	-2.05	.04
<i>Random effects</i>	<i>Variance</i>	<i>SD</i>				<i>Variance</i>	<i>SD</i>				<i>Variance</i>	<i>SD</i>				<i>Variance</i>	<i>SD</i>		
Item intercept	0.01	0.10				0.01	0.12				0.16	0.40				0.02	0.15		
Age group Item	—	—				—	—				0.00	0.06				—	—		
Phrase type Item	—	—				—	—				—	—				0.00	0.07		
Participant intercept	0.03	0.17				0.04	0.19				0.07	0.26				0.12	0.35		
Age group Participant	—	—				—	—				—	—				0.00	0.04		
Residual	0.18	0.43				0.20	0.44				0.30	0.55				—	—		

Notes: The estimated value of each predictor was rounded and kept to two decimal places. Some random effects were discarded (as shown by “—”) due to model fitting, and the analyses of variance in fixed and random effects showed no significant differences between the fitted model and the original model(s).

Table 4. Summary of selected models (with log-phrase frequency) in the phrase-final word analysis

Fixed effects	First fixation duration					First pass reading time					Dwell time					Fixation count			
	Estimate	SE	df	t	pr	Estimate	SE	df	t	pr	Estimate	SE	df	t	pr	Estimate	SE	z	pr
Intercept	6.26	0.14	66.50	46.29	<.001	6.48	0.16	122.20	40.66	<.001	7.25	0.21	115.50	34.28	<.001	1.84	0.20	9.06	<.001
Age group	-0.11	0.03	69.24	-3.28	<.01	-0.16	0.04	116.50	-4.17	<.001	-0.27	0.05	106.50	-5.53	<.001	-0.25	0.05	-5.49	<.001
log-Phrase frequency	0.03	0.02	68.85	1.45	.15	0.02	0.02	1581.00	1.03	.30	0.02	0.02	1268.00	0.70	.49	0.02	0.03	0.67	.50
Age group × log-Phrase frequency	-0.01	0.01	71.55	-2.14	.04	-0.01	0.00	1803.00	-1.87	.06	-0.01	0.01	1800.00	-1.67	.09	-0.01	0.01	-1.25	.21
Random effects	Variance	SD				Variance	SD				Variance	SD			Variance	SD			
Item intercept	0.03	0.17				0.01	0.12				0.04	0.19			0.08	0.29			
Age group Item	0.00	0.05				—	—				—	—			0.00	0.03			
log-Phrase frequency Item	0.00	0.03				—	—				—	—			—	—			
Participant intercept	0.07	0.27				0.04	0.19				0.07	0.26			0.13	0.35			
Age group Participant	0.01	0.09				—	—				—	—			0.00	0.04			
log-Phrase frequency Participant	0.00	0.01				—	—				—	—			—	—			
Residual	0.18	0.42				0.20	0.44				0.30	0.55			—	—			

Notes: The estimated value of each predictor was rounded and kept to two decimal places. Some random effects were discarded (as shown by “—”) due to model fitting, and the analyses of variance in fixed and random effects showed no significant differences between the fitted model and the original model(s).

across the early, middle, and late measures ($t_{FFD} = 1.18, p > .10; t_{FPRT} = 1.11, p > .10; t_{DT} = 1.43, p > .10; z_{FC} = 1.58, p > .10$). Similarly, *phrase frequency* was not significant across any of the measures ($t_{FFD} = 1.45, p > .10; t_{FPRT} = 1.03, p > .10; t_{DT} = 0.70, p > .10; z_{FC} = 0.67, p > .10$). *Age group* was a significant predictor across all the measures in the models with phrase type ($t_{FFD} = -6.01, p < .001; t_{FPRT} = -6.16, p < .001; t_{DT} = -7.30, p < .001; z_{FC} = -8.77, p < .001$), as well as the models with phrase frequency ($t_{FFD} = -3.28, p < .01; t_{FPRT} = -4.17, p < .001; t_{DT} = -5.53, p < .001; z_{FC} = -5.49, p < .001$).

Post hoc analysis revealed that Grade 4 readers were faster than Grade 3 readers (in models with phrase type: $t_{FFD} = 2.83, p = .02$, effect size = 0.41, $t_{FPRT} = 3.01, p = .01$, effect size = 0.49, $t_{DT} = 2.61, p = .03$, effect size = 0.48, $z_{FC} = 2.20, p = .07$, effect size = 0.17; in models with phrase frequency: $t_{FFD} = 3.02, p = .01$, effect size = 0.42, $t_{FPRT} = 3.01, p = .01$, effect size = 0.49, $t_{DT} = 2.70, p = .02$, effect size = 0.47, $z_{FC} = 2.36, p = .05$, effect size = 0.19). Adults were faster readers than Grade 4 readers (in models with phrase type: $t_{FFD} = 2.63, p = .03$, effect size = 0.36, $t_{FPRT} = 2.60, p = .03$, effect size = 0.39, $t_{DT} = 3.98, p < .001$, effect size = 0.67, $z_{FC} = 5.86, p < .001$, effect size = 0.40; in models with phrase frequency: $t_{FFD} = 2.54, p = .04$, effect size = 0.37, $t_{FPRT} = 2.60, p = .03$, effect size = 0.39, $t_{DT} = 4.06, p < .001$, effect size = 0.67, $z_{FC} = 5.49, p < .001$, effect size = 0.39).

Crucially, we found a significant interaction between *age group* and *phrase type* across the early, middle, and late measures ($t_{FFD} = -1.95, p = .05; t_{FPRT} = -1.94, p = .05; t_{DT} = -2.42, p = .02; z_{FC} = -2.05, p = .04$). The interaction is plotted in Figure 2. Further comparisons showed that adults, but not third or fourth graders, read the final word in collocations significantly faster than in control phrases across the four measures (adults: $t_{FFD} = 2.60, p = .01$, effect size = 0.24, $t_{FPRT} = 2.60, p = .01$, effect size = 0.25, $t_{DT} = 3.15, p < .01$, effect size = 0.31, $z_{FC} = 2.11, p = .03$, effect size = 0.15; Grade 3: $t_{FFD} = 0.28, p > .10$, effect size = 0.03, $t_{FPRT} = 0.37, p > .10$, effect size = 0.04, $t_{DT} = 0.19, p > .10$, effect size = 0.03, $z_{FC} = -0.11, p > .10$, effect size < 0.01; Grade 4: $t_{FFD} = 1.67, p = .10$, effect size = 0.18, $t_{FPRT} = 1.85, p = .07$, effect size = 0.20, $t_{DT} = 1.24, p > .10$, effect size = 0.17, $z_{FC} = 0.66, p > .10$, effect size = 0.05).

The interaction between *age group* and *phrase frequency* was found significant in the early measure ($t_{FFD} = -2.14, p = .04$), and marginally significant in the middle and dwell time measure ($t_{FPRT} = -1.87, p = .06; t_{DT} = -1.67, p = .09$), but was not significant in the fixation count measure ($z_{FC} = -1.25, p > .10$). The interaction is plotted in Figure 3. Further analysis of the three eye-tracking measures showed that adults read the final word in higher frequency phrases faster than in lower frequency sequences ($F_{FFD} = 9.98, p < .01$, effect size = 0.13; $F_{FPRT} = 9.96, p < .01$, effect size = 0.13; $F_{DT} = 16.10, p < .001$, effect size = 0.20). Crucially, this processing advantage emerged in the early and middle measures in fourth graders ($F_{FFD} = 3.85, p = .05$, effect size = 0.01; $F_{FPRT} = 3.97, p = .05$, effect size = 0.01), but not in the late measures ($F_{DT} = 1.54, p > .10$, effect size < 0.01). No processing advantage for the final word in higher frequency phrases versus lower frequency ones was found in third graders ($F_{FFD} = 0.11, p > .10$, effect size < 0.01; $F_{FPRT} = 0.18, p > .10$, effect size < 0.01; $F_{DT} = 1.25, p > .10$, effect size < 0.01).

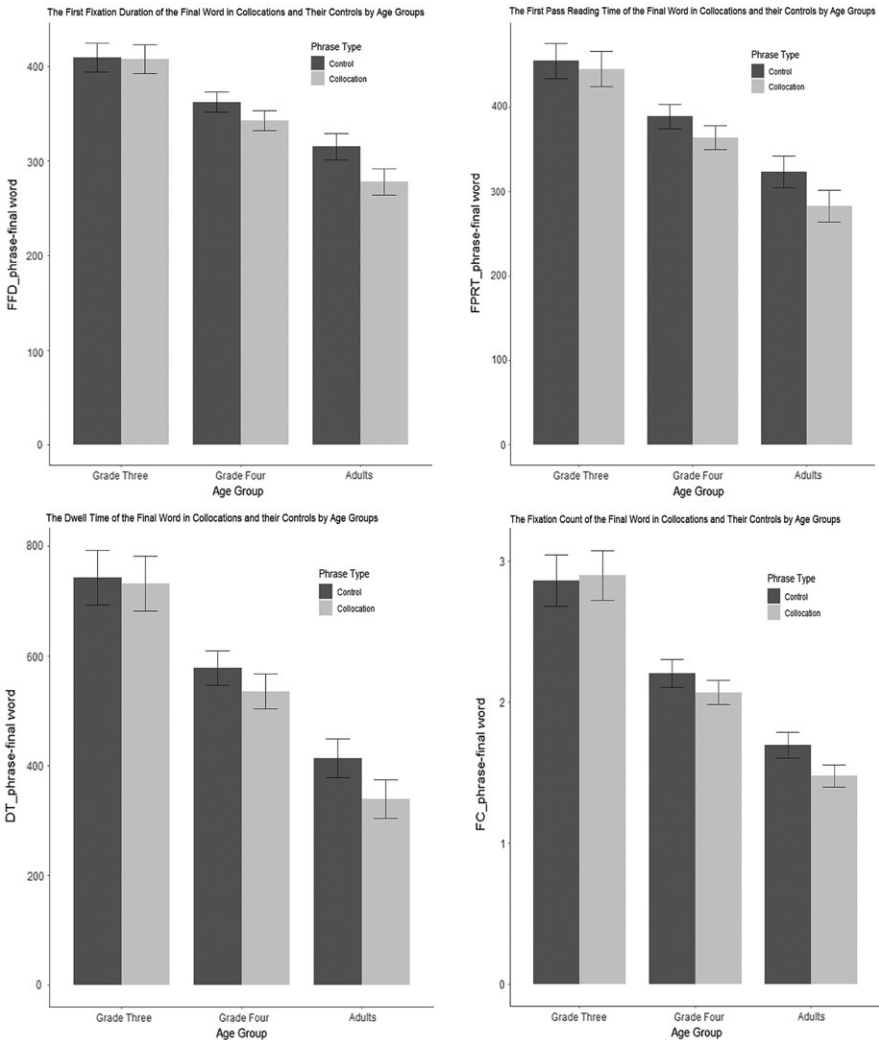


Figure 2. The interaction between age group and phrase type for the first fixation duration, first pass reading time, dwell time, and fixation count in the phrase-final word analysis, with 95% confidence intervals.

Analysis of the whole phrase

We further analyzed the reading of the whole phrase (Table 5: models with phrase type; Table 6: models with phrase frequency). Both *phrase type* and *phrase frequency* were significant predictors across the dwell time measure, but not other eye-tracking measures (in models with phrase type: $t_{FFD} = -0.15, p > .10$; $t_{FPRT} = 1.59, p > .10$; $t_{DT} = 3.31, p < .001$; $z_{FC} = 1.63, p = .10$; in models with phrase frequency: $t_{FFD} = -0.15, p > .10$; $t_{FPRT} = 1.47, p > .10$; $t_{DT} = 2.62, p < .01$; $z_{FC} = 1.06, p > .10$). Further analyses of the dwell time measure showed that collocations were overall

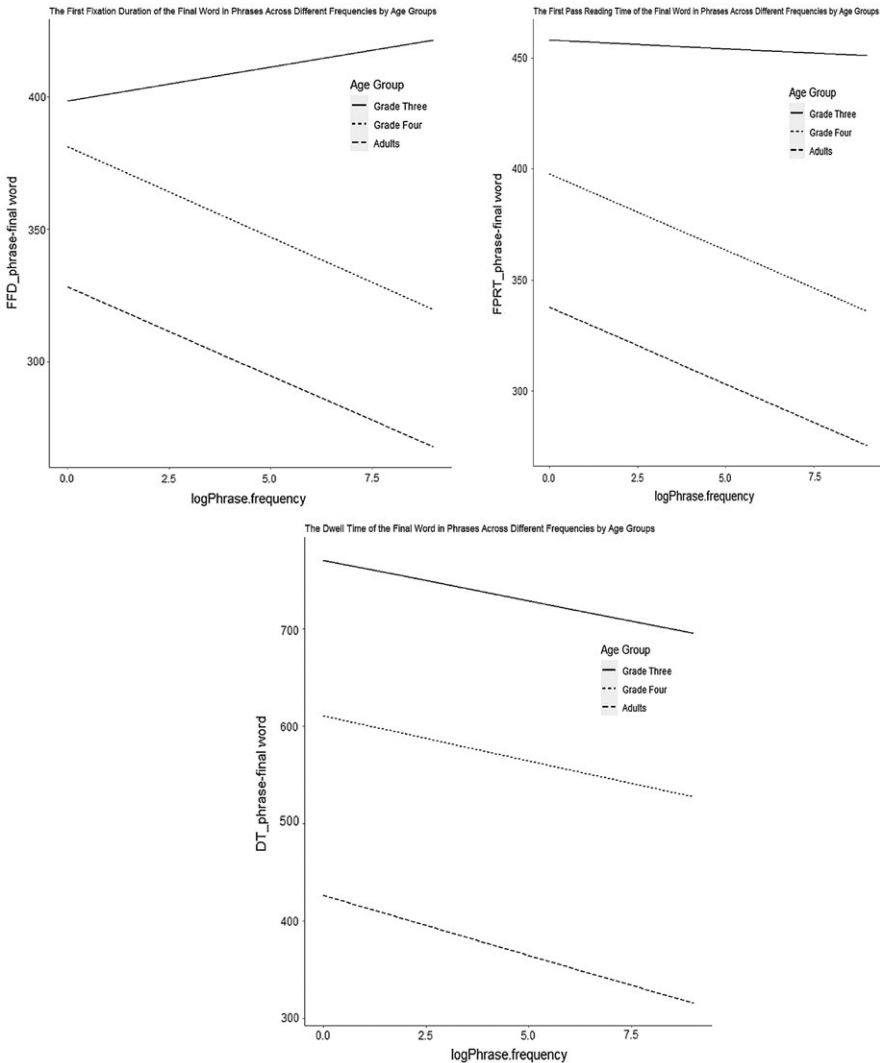


Figure 3. The interaction between age group and log-phrase frequency for the first fixation duration, first pass reading time, and dwell time in the phrase-final word analysis.

processed faster than their controls ($t_{DT} = 4.65, p < .001$, effect size = 0.46), and that higher frequency phrases were read faster than lower frequency ones ($F_{DT} = 45.27, p < .001$, effect size = 0.21). Age group was also found significant in the middle and late measures (in models with phrase type: $t_{FFD} = -0.68, p > .10$; $t_{FPRT} = -4.02, p < .001$; $t_{DT} = -8.28, p < .001$; $z_{FC} = -9.50, p < .001$; in models with phrase frequency: $t_{FFD} = -0.41, p > .10$; $t_{FPRT} = -2.35, p = .02$; $t_{DT} = -5.50, p < .001$; $z_{FC} = -6.77, p < .001$). Post hoc analyses of the three eye-tracking measures revealed that although Grade 4 readers were not significantly faster than Grade 3 readers (in models with phrase type: $t_{FPRT} = 1.85, p > .10$, effect size = 0.31, $t_{DT} = 2.30, p = .06$,

Table 5. Summary of selected models (with phrase type) in the whole-phrase analysis

<i>Fixed effects</i>	First fixation duration					First pass reading time					Dwell time					Fixation count			
	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>pr</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>pr</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>pr</i>	<i>Estimate</i>	<i>SE</i>	<i>z</i>	<i>pr</i>
Intercept	5.46	0.09	62.87	59.42	<.001	6.68	0.20	59.90	33.60	<.001	8.38	0.20	60.29	42.88	<.001	3.04	0.17	17.85	<.001
Age group	-0.01	0.02	61.59	-0.68	.50	-0.19	0.05	59.12	-4.02	<.001	-0.38	0.05	58.92	-8.28	<.001	-0.38	0.04	-9.50	<.001
Phrase type	-0.01	0.05	56.12	-0.15	.88	0.11	0.07	1847.46	1.59	.11	0.18	0.06	1233.07	3.31	<.001	0.11	0.07	1.63	.10
Age group × Phrase type	0.00	0.01	56.31	-0.12	.91	-0.04	0.02	1953.60	-2.22	.03	-0.07	0.01	1950.37	-5.80	<.001	-0.04	0.01	-2.39	.02
<i>Random effects</i>	<i>Variance</i>		<i>SD</i>		<i>Variance</i>		<i>SD</i>		<i>Variance</i>		<i>SD</i>		<i>Variance</i>		<i>SD</i>				
Item intercept	0.03	0.16			0.02	0.13			0.03	0.17			0.12	0.35					
Age group Item	0.00	0.03			—	—			—	—			0.00	0.07					
Phrase type Item	—	—			—	—			—	—			0.02	0.14					
Participant intercept	0.02	0.13			0.09	0.29			0.09	0.30			0.06	0.25					
Phrase type Participant	0.00	0.04			—	—			—	—			—	—					
Residual	0.10	0.31			0.41	0.64			0.22	0.47			—	—					

Notes: The estimated value of each predictor was rounded and kept to two decimal places. Some random effects were discarded (as shown by “—”) due to model fitting, and the analyses of variance in fixed and random effects showed no significant differences between the fitted model and the original model(s).

Table 6. Summary of selected models (with log-phrase frequency) in the whole-phrase analysis

Fixed effects	First fixation duration					First pass reading time					Dwell time					Fixation count			
	Estimate	SE	df	t	pr	Estimate	SE	df	t	pr	Estimate	SE	df	t	pr	Estimate	SE	z	pr
Intercept	5.47	0.11	107.80	48.09	<.001	6.50	0.23	113.70	27.80	<.001	8.13	0.22	92.38	37.35	<.001	2.94	0.21	13.94	<.001
Age group	-0.11	0.03	110.20	-0.41	.68	-0.13	0.06	109.80	-2.35	.02	-0.28	0.05	85.72	-5.50	<.001	-0.34	0.05	-6.77	<.001
log-Phrase frequency	0.00	0.01	78.08	-0.15	.88	0.04	0.03	1841.00	1.47	.14	0.06	0.02	1191.00	2.62	<.01	0.02	0.02	1.06	.29
Age group × log-Phrase frequency	0.00	0.00	94.32	-0.20	.84	-0.01	0.01	1951.00	-2.09	.04	-0.02	0.00	1949.00	-4.92	<.001	-0.01	0.00	-1.83	.07
Random effects	Variance	SD					Variance	SD			Variance	SD			Variance	SD			
Item intercept	0.02	0.16					0.02	0.13			0.03	0.18			0.03	0.16			
Age group Item	0.00	0.03					—	—			—	—			—	—			
Participant intercept	0.02	0.13					0.09	0.29			0.09	0.30			0.05	0.21			
Age group Participant	—	—					—	—			—	—			0.00	0.01			
Residual	0.10	0.32					0.41	0.64			0.22	0.47			-	-			

Notes: The estimated value of each predictor was rounded and kept to two decimal places. Some random effects were discarded (as shown by “—”) due to model fitting, and the analyses of variance in fixed and random effects showed no significant differences between the fitted model and the original model(s).

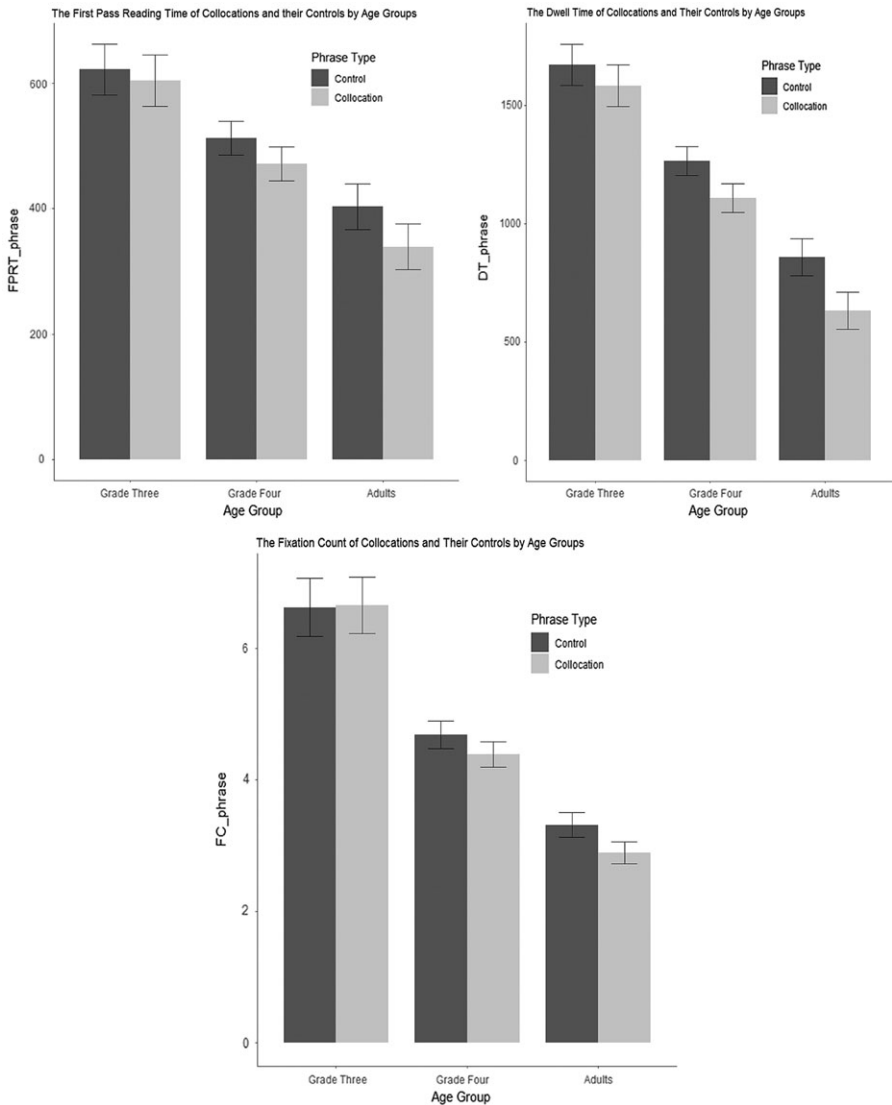


Figure 4. The interaction between age group and phrase type for the first pass reading time, dwell time, and fixation count in the whole phrase analysis, with 95% confidence intervals.

effect size = 0.50, $z_{FC} = 0.72$, $p > .10$, effect size = 0.05; in models with phrase frequency: $t_{FPRT} = 1.85$, $p > .10$, effect size = 0.31, $t_{DT} = 2.30$, $p = .06$, effect size = 0.50, $z_{FC} = 0.78$, $p > .10$, effect size = 0.05), adults read faster than Grade 4 in late measures (in models with phrase type: $t_{FPRT} = 1.81$, $p > .10$, effect size = 0.28, $t_{DT} = 5.37$, $p < .001$, effect size = 1.10, $z_{FC} = 9.90$, $p < .001$, effect size = 0.69; in models with phrase frequency: $t_{FPRT} = 1.81$, $p > .10$, effect size = 0.28, $t_{DT} = 5.37$, $p < .001$, effect size = 1.10, $z_{FC} = 10.27$, $p < .001$, effect size = 0.70).

The interaction between *age group* and *phrase type* was found significant in the middle and late measures ($t_{FPRT} = -2.22, p = .03; t_{DT} = -5.80, p < .001; z_{FC} = -2.39, p = .02$), but not in the early measure ($t_{FFD} = -0.12, p > .10$). The interaction is plotted in Figure 4. Further comparisons showed that adults read collocations faster than controls ($t_{FPRT} = 2.45, p = .02$, effect size = 0.20; $t_{DT} = 6.58, p < .001$, effect size = 0.73; $z_{FC} = 3.11, p < .01$, effect size = 0.15). Critically, Grade 4 read collocations faster than controls in the late measure ($t_{FPRT} = 1.56, p > .10$, effect size = 0.15; $t_{DT} = 3.27, p < .01$, effect size = 0.40; $z_{FC} = 0.68, p > .10$, effect size = 0.04). No such processing advantage was found in Grade 3 readers ($t_{FPRT} = -0.34, p > .10$, effect size < 0.01; $t_{DT} = 1.12, p > .10$, effect size = 0.13; $z_{FC} = 0.12, p > .10$, effect size < 0.01).

Finally, the interaction between *age group* and *phrase frequency* was found significant in middle and late measures ($t_{FPRT} = -2.09, p = .04; t_{DT} = -4.92, p < .001$), approaching significance in the fixation count measure ($z_{FC} = -1.83, p = .07$), and not significant in the early measure ($t_{FFD} = -0.20, p > .10$). The interaction is plotted in Figure 5. Further analysis revealed that adults read higher frequency phrases faster than lower frequency phrases ($F_{FPRT} = 9.45, p < .01$, effect size = 0.11; $F_{DT} = 75.92, p < .001$, effect size = 0.79; $F_{FC} = 12.84, p < .001$, effect size = 0.14). Of importance, Grade 4 readers read higher frequency phrases faster than lower frequency phrases, as evident in the dwell time measure ($F_{FPRT} = 2.78, p = .09$, effect size = 0.01; $F_{DT} = 11.91, p < .001$, effect size = 0.21; $F_{FC} = 1.60, p > .10$, effect size < 0.01). No such processing advantage was found in Grade 3 readers ($F_{FPRT} = 0.07, p > .10$, effect size < 0.01; $F_{DT} = 1.85, p > .10$, effect size < 0.01; $F_{FC} = 0.49, p > .10$, effect size < 0.01).

Discussion

The processing advantage for a wide range of MWEs (e.g., collocations, idioms, lexical bundles, etc.) over infrequent control phrases has long been attested in the literature. However, the relevant evidence pertains almost exclusively to MWE processing in adults, in most cases, educated university students. Whether or not other populations and age groups are sensitive to phrase frequency effects during reading has, to date, been largely disregarded. In the present investigation, we sought to address this gap by asking 8- and 9-year-old primary school pupils, as well as a group of adults (L1 Mandarin), to read phrases varying in frequency embedded in sentence context. In particular, we were interested in how phrase type, phrase frequency, and age might influence participants' reading of the phrase-final word and the whole phrase. The following findings emerged.

We found significant main effects of *phrase type* and *phrase frequency*. In particular, we found that, overall, collocations were read faster than control phrases, and, correspondingly, higher frequency phrases were read faster than lower frequency ones. Both the phrase type and phrase frequency effects were observed in the total reading time measure (dwell time). These findings are in line with earlier research involving adults, pointing to an important role of frequency in language acquisition, processing, and use.

Further, *age* was found to be a significant predictor of (general) reading times across the analyses and eye-tracking measures, with the youngest (Grade 3) readers being the slowest, the oldest (adult) readers being the fastest, and Grade 4 readers

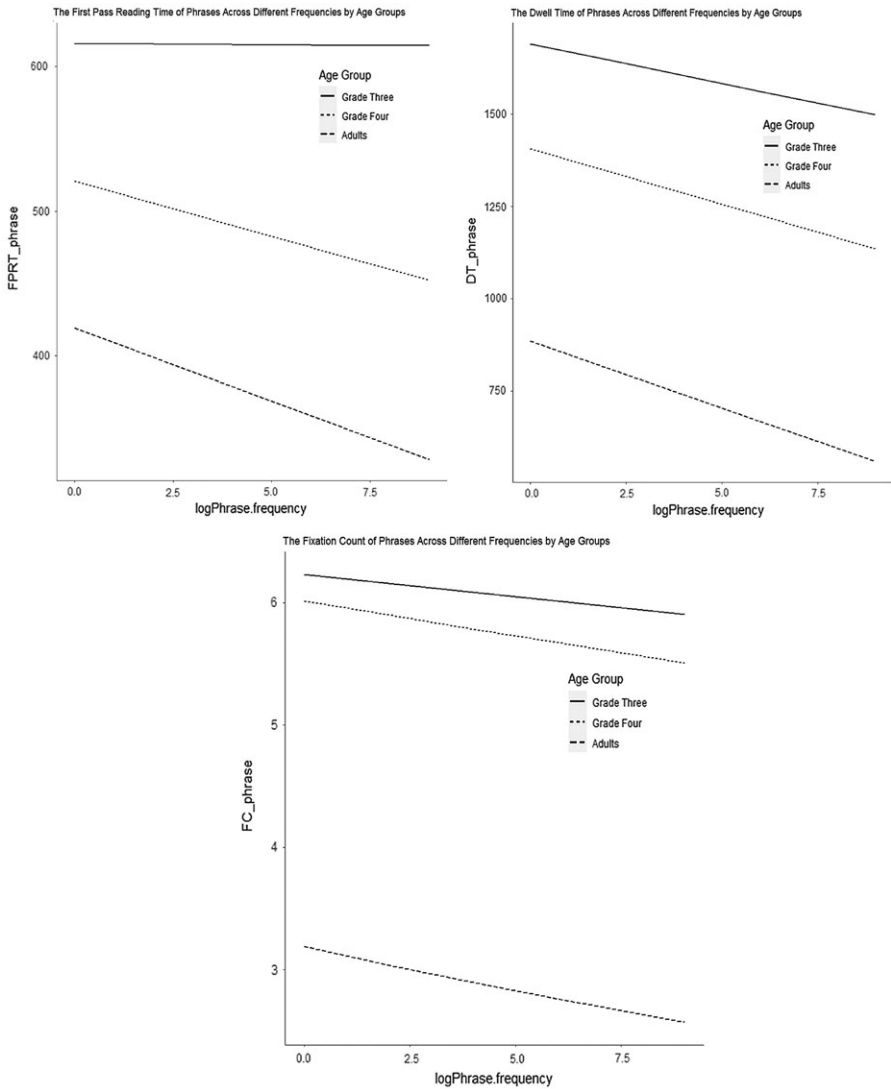


Figure 5. The interaction between age group and log-phrasal frequency for the first pass reading time, dwell time, and fixation count in the whole phrase analysis.

being in the middle. The incremental reading pattern as a function of age supports the studies on the development of reading fluency in readers of various ages (e.g., Landerl & Wimmer, 2008).

The analyses of the two interest areas pointed to significant two-way interactions involving age, phrase type, and phrase frequency. First, age was found to interact with phrase type in the analysis of the phrase-final word, such that the noun in collocations was read faster than the noun in control phrases by adults, but not third or fourth graders. Second, age interacted with phrase frequency, revealing faster

reading for the final word in higher frequency versus lower frequency phrases in adults, and, critically, in fourth graders. While the effect for the adult readers was robust and was evident across the board, the effect found for Grade 4 was relatively small ($p = .05$) and was observed in the early and middle, but not late measures. Further, age interacted both with phrase type and phrase frequency in the whole phrase analysis. The interactions were evident across the middle and late measures, but were, in particular, evident in the total reading time (dwell time) measure. As expected, adults read collocations and higher frequency phrases consistently faster than controls and lower frequency sequences, respectively. Grade 4 readers also read collocations and higher frequency phrases reliably faster than controls and lower frequency phrases, respectively. Although phrase type and phrase frequency effects for Grade 4 readers emerged only in the analysis of one late measure, dwell time, the effects were strong (phrase type: $p < .01$; phrase frequency: $p < .001$).

The results of the present study are the first in the literature to point to an emerging effect of phrase frequency in primary school children during reading. While adults have consistently been shown to be attuned to phrase frequency in language production and comprehension (e.g., Arnon & Snider, 2010; Kapatsinski & Radicke, 2009; Siyanova-Chanturia, Conling, van Hauven, 2011; Siyanova-Chanturia & Janssen, 2018; Tremblay & Baayen, 2010; Tremblay et al., 2011), very little evidence exists with regard to children. Bannard and Matthews (2008) demonstrated that children as young as 3 were sensitive to phrase frequency distributions during elicited language production (repetition). Reading, however, is a more complex cognitive process requiring the decoding of symbols, that is, “translating signs and symbols into meanings” (Robeck & Wallace, 2018, p. 27). During reading, “literal and higher-level comprehension processes occur as written discourse is decoded” (Robeck & Wallace, 2018, p. 24). Reading, thus, encompasses both decoding and comprehension; shifting between the two can be challenging for beginner readers. Reading Chinese characters can be even more challenging, given their number⁴ and “visual complexity” (Liu et al., 2015, p. 307). It is thus, perhaps, not surprising that only Grade 4 (relatively skilled), and not Grade 3 (less skilled), readers exhibited emerging sensitivity to phrase frequency and phrase type during reading. It is also clear, however, that further studies are needed, in the domain of both production and comprehension, and with children of various age groups and from different L1 backgrounds.

In Bannard and Matthews’s study (2008), children’s age was not found to interact with phrase frequency, which was taken to show “continuity in frequency effects across development” (p. 247). The findings of our reading study point to a different developmental picture, as age interacted with both phrase type and phrase frequency, across the analyses of the two interest areas and a range of early, middle, and late eye-tracking measures. This underlines the need for a variety of measures and paradigms to be employed in order to tap into the processes associated with the development of phrase frequency effects in language comprehension. Our findings suggest that children need sufficient experience with reading and exposure to written characters to be able to start to exhibit sensitivity to phrase frequency during naturalistic reading, and that the age at which this may happen (at least for L1 Chinese) is around nine. This tenet, however, is speculative and is in need of further exploration involving younger readers, for example first and second graders.

(Note that prior to the experiment, the third and fourth graders would have been reading in Chinese for only about 2 and 3 years, respectively).

From a theoretical standpoint, and in line with previous research (e.g., Arnon & Cohen Priva, 2013, 2014; Bybee & Scheibman, 1999; Siyanova-Chanturia, Conklin, & van Heuven, 2011), we take both adult and child readers' sensitivity to the distributional properties of MWEs as being problematic for the words-and-rules view of language. This approach views the lexicon (memorized and stored forms) and grammar (a collection of rules) as distinct entities (e.g., Pinker, 1998, 1999; Pinker & Ullman, 2002). According to this approach, frequency should play a role in the processing of memorized forms, such as morphologically simple forms and idiosyncratic phrases (e.g., idioms). Compositional phrases, it is argued, are necessarily computed and hence should not be subject to phrase frequency effects. The results of our study are at odds with this view. On the contrary, our findings are compatible with the emergentist accounts of language acquisition, processing, and use. These accounts encompass a variety of approaches, such as connectionist models (e.g., Christiansen & Chater, 1999; Elman, 1990; Monaghan, Chang, Welbourne, & Brysbaert, 2017; Rumelhart & McClelland, 1986), usage-based theories (e.g., Bybee, 1998; Goldberg, 2006; Langacker, 1987; Tomasello, 2003), and exemplar-based models of language processing (e.g., Abbot-Smith & Tomasello, 2006; Bod, 2006; Pierrehumbert, 2001). These models are not uniform and vary in several ways. However, a key tenet that these theories share is an important role of frequency in the processing of any and all linguistic forms (e.g., memorized and stored forms, such as morphologically simple words and idioms, as well as morphologically complex words, multiword phrases, and longer stretches of language).

According to the emergentist accounts, the frequency with which linguistic exemplars varying in size, complexity, and level of abstractness occur is a decisive factor in what language users learn and represent. Consistent with connectionist, usage-based and exemplar-based approaches, we take the phrase frequency effects observed in adults and, to a lesser extent, children to reflect language users' sensitivity to the distribution of linguistic information at various grain-sizes. We expand the current body of evidence (e.g., Arnon & Cohen Priva, 2013, 2014; Bannard & Matthews, 2008; Janssen & Barber, 2012; Siyanova-Chanturia, Conklin, & van Heuven, 2011) by showing that by the age of 9, children are already attuned to higher versus lower frequency phrases during naturalistic reading. Because the younger counterparts were not able to reliably differentiate between collocations versus controls, we argue that unlike language production (Bannard & Matthews, 2008) and, possibly, aural comprehension, phrase frequency effects may be relatively late to manifest themselves during reading. While 9-year-old children would have had years of experience speaking and hearing the language, they would have far more limited experience reading in their L1.

It is noteworthy that in the present study, phrase frequency was conceptualized and tested as a categorical (phrase type) and a continuous (phrase frequency) variable. It has been argued that a continuous measure of frequency, rather than a binary one (low vs. high), may be a better predictor of response times (e.g., Arnon & Snider, 2010). However, because the two are distinct in what they represent and what they reflect in processing terms, we decided to look at both. For example, Siyanova-Chanturia, Conklin, and van Heuven (2011) and Siyanova-Chanturia

and Janssen (2018) employed both measures of frequency and found some noteworthy differences. In Siyanova-Chanturia, Conklin, and van Heuven (2011), native speakers and proficient second-language speakers read English binomials (*time and money*) faster than reversed forms (*money and time*). On the contrary, less proficient learners did not exhibit such an advantage. When phrase frequency was treated as a continuous variable, *all* participants showed phrase frequency effects. Using the same experimental materials, Siyanova-Chanturia and Janssen (2018) had native and L2 speakers perform a phrase-elicitation (production) task. Native speakers' articulatory durations were affected by phrase frequency, but not phrase type. Learners' durations were not influenced either by phrase frequency or phrase type. Participants in Siyanova-Chanturia, Conklin, and van Heuven (2011) and Siyanova-Chanturia and Janssen (2018) were more likely to be affected by phrase frequency than phrase type. The current findings pertaining to adults suggest that these readers were similarly attuned to both measures of frequency, categorical and continuous, irrespective of the target interest area(s). In our study, however, the processing pattern for Grade 4 readers was affected more by the continuous measure of frequency (phrase frequency) than the categorical one (phrase type). First, fourth graders read higher frequency phrases faster than lower frequency ones in the phrase-final word analysis, while no such processing advantage was found for collocations over controls (i.e., phrase type). Second, the magnitude of the continuous phrase frequency effect ($p < .001$) for these readers in the whole phrase analysis was greater than that of phrase type ($p < .01$). Our findings, thus, lend support to Arnon and Snider (2010), suggesting that the continuous measure of frequency may be a more powerful and accurate predictor of reading behavior than the binary one, with the frequency effect being more readily detectable. This may be, particularly, relevant where participants are not yet fluent or experienced readers (e.g., young L1 readers, or L2 speakers).

The findings of the present investigation also contribute to the current body of knowledge specific to frequency effects in Chinese. Previous research, albeit solely with adults, showed that Chinese readers are sensitive both to word and character frequency (e.g., Rayner, Li, Juhasz, & Yan, 2005; Yan, Tian, Bai, & Rayner, 2006; Zhang & Peng, 1992). More recently, Kong et al. (2016) and Yi et al. (2017) extended these results to phrase frequency effects in language processing. Our findings provide further evidence that Chinese adult readers are attuned to the distribution of linguistic information at various grain-sizes, that is, not just characters and words as has previously been shown (e.g., Yan, et al., 2006; Zhang & Peng, 1992) but also sequences above the word level.

In addition, our study extends the current literature by showing that, not unlike adults, primary school children are also attuned to phrase frequency during naturalistic reading. The fact that older and more skilled fourth graders, but not younger and less skilled third graders, read higher frequency sequences faster than lower frequency ones reflects the important role of exposure to language and general experience with reading. It is a well-documented finding that more proficient language users are more likely to exhibit phrase frequency effects than less proficient speakers (e.g., Siyanova-Chanturia, Conklin, & van Heuven, 2011; Siyanova-Chanturia & Janssen, 2018; Vilkaite & Schmitt, 2017). The findings of the present investigation suggest that our Chinese third graders (who were not yet skilled readers) would not

have had enough experience with reading language to start to exhibit facilitative processing for collocations over controls, or higher frequency phrases over lower frequency ones. That no processing advantage, even marginal, was observed in any of the analyses performed (two areas of interest; four eye-movement measures), further adds credibility to this finding. In addition, even the data for Grade 4 readers point to *incipient* phrase frequency effects, as they were largely observed in one late measure. Thus, we can conclude that even these more proficient and skilled (Grade 4) readers would not have had enough experience with written language to exhibit phrase frequency effects across the board, as was the case with the adult readers.

From a more methodological standpoint, our study is a testament to how the use of the eye-tracking methodology can help us forge a fuller understanding of the mechanisms behind MWE online processing. It is interesting to note how the use of a variety of eye-tracking measures, along with different interest areas (phrase-final word and whole phrase), allowed for a detailed picture of different reading patterns to emerge. While adults comfortably read collocations and higher frequency phrases faster than controls and lower frequency phrases, respectively, across the analyses of both interest areas, this processing advantage was largely confined to the whole-phrase analysis for the less fluent and less experienced fourth graders. This again underscores the necessity for a variety of interest areas to be included in MWE research. Remarkably, we found no main effects of either phrase type or phrase frequency in the analysis of the phrase-final word. When the whole phrase was considered, both frequency effects (phrase type and phrase frequency) were highly significant in the late, but never in the early or middle, measures. This lends support to Siyanova-Chanturia, Conklin, and Schmitt (2011) who argue that late measures may be more sensitive to potential differences between MWEs and novel speech than early and middle measures.

In conclusion, the results of the present reading investigation have reaffirmed phrase frequency effects in adults, and, critically, have pointed to emerging phrase frequency effects in primary school children. The use of the eye-tracking methodology has further allowed us to tap into the mechanisms associated with phrasal processing in children and adults, and the differences between phrase frequency conceptualized as a dichotomy versus a continuum. While the finding of incipient phrase frequency effects in children is a novel and important one, future research should replicate and expand on this finding in a number of ways. First, it is desirable to focus on children of a wider age range (e.g., 8-year-olds to early teens), in order to paint a more detailed picture of the development of frequency effects as a function of the ever-changing exposure to the target structures and greater experience with reading, in general. Second, reading studies should endeavor to incorporate a range of language backgrounds (e.g., alphabetic, syllabic, and logographic), as languages are likely to differ in how reading development happens in childhood (e.g., Chinese primary school children are first exposed to pinyin, before being taught to read in simplified Chinese characters; e.g., Chen, 2016), and a variety of target sequences (e.g., different phrasal combinations, such as V+N, Adj+N, etc., have been shown to carry different learning, and, possibly, processing burden; e.g., Peters, 2016; Wolter & Yamashita, 2014). Third and finally, given the scarcity of studies looking at phrase frequency effects in children, it is important for various modalities to be explored, targeting both production and comprehension. While adults, having

generally experienced huge amounts of exposure with the target language, have been shown to be attuned to phrase frequency in comprehension (i.e., reading and listening; e.g., Arnon & Snider, 2010; Hernández, Costa, & Arnon, 2016; Siyanova-Chanturia, Conklin, & van Heuven, 2011), as well as production (i.e., repetition/speaking; e.g., Bell et al., 2003; Bybee & Scheibman, 1999; Janssen & Barber, 2012; Siyanova-Chanturia & Janssen, 2018; Tremblay & Tucker, 2011), children's behavior is likely to differ depending on the modality and the varied experience they have had using the language (e.g., a 9-year-old would have had years of experience listening to and speaking in the L1, but only 2 to 3 years engaging in such a complex cognitive activity as reading). Finally, while the MWEs used in the present study were extracted from an adult reference corpus, where possible, future studies should endeavor to use corpora of child speech and writing (as was done in Bannard & Matthews, 2008).

In sum, although the results of the present study require further interrogation and replication, we hope to have contributed to the existing body of knowledge pertinent to phrase frequency effects and, in particular, to have paved the way for future studies into MWE processing in children.

Acknowledgments. This research was supported by the Fundamental Research Fund for the Central Universities and the Research Fund of Beijing Language and Culture University (No. 15YCX152) to the first author, and the Major Project of National Social Science Foundation of China (No.17ZDA305) to the second author. We would like to thank the students from the Third Experimental Primary School in Haidian District and the Primary School Affiliated to the Beijing Petroleum Institute for their participation in the study.

Notes

1. The BCC corpus contains around 15 billion characters and is “ideal as a data source for studies in linguistics as well as applied linguistics” (Xun, Rao, Xiao, & Zang, 2016, p. 118). The corpus contains texts from newspapers, literature, Weibo (spoken texts), and technology. This corpus is believed to be the largest Chinese corpus to date, and one of the “three most widely used Chinese general-purpose corpora” (Xu, 2015, p. 219).
2. We checked word familiarity with the child participants after the experiment. No unknown words or characters were reported by any of the participants.
3. The test included 100 Chinese characters, in descending order of their frequency. These characters were retrieved from the top 3,000 most frequently used characters appearing on the character frequency list in modern Chinese corpus (State Language Commission, 2012). These 3,000 characters (in descending order of frequency) were first divided into 100 groups, with 30 characters in each group. One character was then randomly selected from each group. The child participants were asked to read out these characters and to provide their meaning(s). One score was obtained for the correct pronunciation and meaning(s).
4. It is estimated that a college graduate knows between 4,000 and 5,000 characters, and between 40,000 and 60,000 words (e.g., DeFrancis et al., 1968; also see Hue, 2003). According to Shu, Chen, Anderson, Wu, and Xuan (2003), over 400 characters are introduced to children in Grade 1 (around 6–7 years of age) and another 700 in Grade 2.

References

- Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *Linguistics Review*, 23, 275–290.
- Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, 56, 349–371.

- Arnon, I., & Cohen Priva, U. (2014). Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent phrases. *Mental Lexicon*, *9*, 377–400.
- Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.
- Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, *10*, 340–357.
- Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning—The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. Ithaca, NY: Cornell University Library. Retrieved from <http://arxiv.org/abs/1506.04967>
- Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge: Cambridge University Press.
- Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gegory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, *113*, 1001–1024.
- BLCU Corpus Center. (2016). Retrieved on July 7, 2018, from BCC, <http://bcc.blcu.edu.cn>
- Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *Linguistics Review*, *23*, 1–30.
- Bolinger, D. (1975). *Aspects of language* (2nd ed.). New York: Harcourt Brace Jovanovich.
- Bretherton, I., McNew, S., Snyder, L. E., & Bates, E. (1983). Individual differences at 20 months: Analytic and holistic strategies in language acquisition. *Journal of Child Language*, *10*, 293–320.
- Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53), New York: Wiley.
- Bybee, J. (1998). The emergent lexicon. *Chicago Linguistics Society*, *34*, 421–435.
- Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of Don't in English. *Linguistics*, *37*, 575–596.
- Carrol, G., & Conklin, K. (2014). Eye-tracking multi-word units: Some methodological questions. *Journal of Eye Movement Research*, *7*, 1–11.
- Carrol, G., & Conklin, K. (2015). Cross language lexical priming extends to formulaic units: Evidence from eye-tracking suggests that this idea “has legs.” *Bilingualism: Language and Cognition*, *20*, 299–317.
- Carrol, G., & Conklin, K. (2019). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*. Advance online publication. doi: [10.1177/0023830918823230](https://doi.org/10.1177/0023830918823230)
- Carrol, G., Conklin, K., & Gyllstad, H. (2016). Found in translation: The influence of the L1 on the reading of idioms in a L2. *Studies in Second Language Acquisition*, *38*, 403–443.
- Chao, Y. R. (1976). *Aspects of Chinese sociolinguistics: Essays*. Stanford, CA: Stanford University Press.
- Chen, L. L. (2016). Hanyu pinyin. In S-W. Chan (Ed.), *The Routledge encyclopaedia of the Chinese language* (pp. 484–504). New York: Routledge.
- Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205.
- Chung, K. K. H., & McBride-Chang, C. (2011). Executive functioning skills uniquely predict Chinese word reading. *Journal of Educational Psychology*, *103*, 909–921.
- Clark, R. (1974). Performing without competence. *Journal of Child Language*, *1*, 1–10.
- Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, *32*, 453–467.
- Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. New York: Cambridge University Press.
- Cruttenden, A. (1981). Item-learning and system-learning. *Journal of Psycholinguistic Research*, *10*, 79–88.
- Defrancis, J., Yung Teng, C. Y., & Yung, C. S. (1968). *Advanced Chinese reader*. New Haven, CT: Yale University Press.

- Duyck, W., van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 663–649.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.
- Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
- Gu, J., & Li, X. (2015). The effects of character transposition within and across words in Chinese reading. *Attention, Perception, & Psychophysics*, *77*, 272–281.
- Hernández, M., Costa, A., & Arnon, I. (2016). More than words: Multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience*, *31*, 785–800.
- Hue, C. W. (2003). Number of characters a college student knows. *Journal of Chinese Linguistics*, *31*, 300–339.
- Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PLOS ONE*, *7*, e3302. doi: [10.1371/journal.pone.0033202](https://doi.org/10.1371/journal.pone.0033202)
- Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal*, *91*, 433–445.
- Kapatsinski, V., & Radicke, J. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In R. Corrigan, E. Moravcsik, H. Ouali, & K. Wheatley (Eds.), *Formulaic language* (pp. 499–522). Amsterdam, Netherlands: Benjamins.
- Kong, L., Zhang, J. X., & Zhang, Y. (2016). Are Chinese correlative conjunctions psychologically real? An investigation of the combination frequency effect. *Psychological Reports*, *119*, 106–123.
- Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology*, *100*, 150–161.
- Langacker, R. W. (1987). *Foundations of cognitive grammar: Vol. 1. Theoretical prerequisites*. Stanford, CA: Stanford University Press.
- Li, W., Anderson, R. C., Nagy, W., & Zhang, H. (2002). Facets of metalinguistic awareness that contribute to Chinese literacy. In W. Li, J. S. Gaffney, & J. L. Packard (Eds.), *Chinese children's reading acquisition: Theoretical and pedagogical issues* (pp. 87–106). Boston: Kluwer Academic.
- Li, T., Wang, Y., Tong, X., & McBride, C. (2017). A developmental study of Chinese children's word and character reading. *Journal of Psycholinguist Research*, *46*, 141–155.
- Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*, 333–370.
- Lieven, E. V. M., Pine, J. M., & Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of Child Language*, *19*, 287–310.
- Liu, D., Chen, X., & Chung, K. K. H. (2015). Performance in a visual search task uniquely predicts reading abilities in third-grade Hong Kong Chinese children. *Scientific Studies of Reading*, *19*, 307–324.
- Liu, P.-P., Li, W.-J., Lin, N., & Li, X.-S. (2013). Do Chinese readers follow the national standard rules for word segmentation during reading? *PLOS ONE*, *8*, e55440. doi: [10.1371/journal.pone.0055440](https://doi.org/10.1371/journal.pone.0055440)
- Locke, J. (1997). A theory of neurolinguistics development. *Brain and Language*, *58*, 265–326.
- Ma, G., & Li, X. (2015). How character complexity modulates eye movement control in Chinese reading. *Reading and Writing*, *28*, 747–761.
- Molinaro, N., & Carreiras, M. (2010). Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biological Psychology*, *83*, 176–190.
- Monaghan, P., Chang, Y. N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1–21.
- Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, *149*, 1–2.
- Packard, J. L. (2000). *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge: Cambridge University Press.
- Packard, J. L. (2016). Chinese psycholinguistics. In S.-W. Chan (Ed.), *The Routledge encyclopaedia of the Chinese language* (pp. 315–327). New York: Routledge.

- Paterson, K. B., Liversedge, S. P., & Underwood, G.** (1999). The influence of focus operators on syntactic processing of short relative clause sentences. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *52A*, 717–737.
- Pellicer-Sánchez, A., & Siyanova-Chanturia, A.** (2018). Eye movements in vocabulary research. *International Journal of Applied Linguistics*, *169*, 5–29.
- Peters, A. M.** (1977). Language learning strategies: Does the whole equal the sum of the parts? *Language*, *53*, 560–573.
- Peters, A. M.** (1983). *Units of language acquisition*. Cambridge: Cambridge University Press.
- Peters, E.** (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, *20*, 113–138.
- Pierrehumbert, J.** (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam, Netherlands: Benjamins.
- Pine, J. M., & Lieven, E. V. M.** (1993). Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language*, *20*, 551–571.
- Pinker, S.** (1998). Words and rules. *Lingua*, *106*, 219–242.
- Pinker, S.** (1999). *Words and rules: The ingredients of language*. New York: Basic Books.
- Pinker, S., & Ullman, M. T.** (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*, 456–463.
- Rayner, K.** (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.
- Rayner, K.** (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*, 1457–1506.
- Rayner, K., Li, X., Juhasz, B. J., & Yan, G.** (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review*, *12*, 1089–1093.
- Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr., C.** (2012). *Psychology of reading* (2nd ed.). New York: Psychology Press.
- Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., & Clifton, C.** (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, *4*, SI21–SI49.
- Robeck, M. C., & Wallace, R. R.** (2018). *The psychology of reading: An interdisciplinary approach* (2nd ed.). New York: Routledge.
- Roberts, L., & Siyanova-Chanturia, A.** (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition*, *35*, 213–235.
- Rumelhart, D. E., & McClelland, J. L.** (1986). *Parallel distributed processing: Vol. 1. Foundations*. Cambridge, MA: MIT Press.
- Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y.** (2003). Properties of school Chinese: Implications for learning to read. *Child Development*, *74*, 27–47.
- Siyanova-Chanturia, A.** (2013). Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *Mental Lexicon*, *8*, 245–268.
- Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., & van Heuven, W. J. B.** (2017). Representation and processing of multi-word expressions in the brain. *Brain and Language*, *175*, 111–122.
- Siyanova-Chanturia, A., Conklin, K., & Schmitt, N.** (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*, 251–272.
- Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B.** (2011). Seeing a phrase “time and again” matters: The role of phrase frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 776–784. doi: [10.1037/a0022531](https://doi.org/10.1037/a0022531)
- Siyanova-Chanturia, A., & Janssen, N.** (2018). Production of familiar phrases: Frequency effects in native speakers and second language learners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*, 2009–2018. doi: [10.1037/xlm0000562](https://doi.org/10.1037/xlm0000562)
- Siyanova-Chanturia, A., & van Lancker Sidtis, D.** (2019). What on-line processing tells us about formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 1–15). London: Routledge.
- Sonbul, S.** (2014). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition*, *18*, 419–437.

- Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language*, *93*, 227–236.
- SR Research Ltd. (2016). Retrieved from <http://www.sr-research.com>
- State Language Commission. (2012). *Chinese character frequency list in modern Chinese corpus*. Retrieved from <http://www.cncorpus.org>.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Tomasello, M., & Brooks, P. J. (1999). Early syntactic development: A construction grammar approach. In M. Barrett (Ed.) *The development of language* (pp. 161–190). Hove, UK: Psychology Press.
- Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioural and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151–173), London: Continuum International Publishing Group.
- Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*, 569–613.
- Tremblay, A., & Tucker, B. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *Mental Lexicon*, *6*, 302–324.
- Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing and Use* (pp. 152–172), Amsterdam, The Netherlands: John Benjamins.
- Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, *22*, 1682–1700.
- Vilkaitė, L. (2016). Are nonadjacent collocations processed faster? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1632–1642.
- Vilkaitė, L., & Schmitt, N. (2017). Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations? *Applied Linguistics*, *40*, 329–354
- Wei, W., Li, X., & Pollatsek, A. (2013). Word properties of a fixated region affect outgoing saccade length in Chinese reading. *Visual Research*, *80*, 1–6.
- Wolter, B., & Yamashita, J. (2014). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, *36*, 1193–1221.
- Wu, X. Y., Anderson, R. C., Li, W. L., Wu, X. C., Li, H., Zhang, J., ... Gaffney, J. S. (2009). Morphological awareness and Chinese children's literacy development: An intervention study. *Scientific Studies of Reading*, *13*, 26–52.
- Xu, J. (2015). Corpus-based Chinese studies: A historical review from the 1920s to the present. *Chinese Language and Discourse*, *6*, 218–244.
- Xun, E., Rao, G., Xiao, X., & Zang, J. (2016). The construction of the BCC corpus in the age of big data. *Corpus Linguistics*, *3*, 93–118.
- 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*, 259–268.
- Yi, W., Lu, S., & Ma, G. (2017). Frequency, contingency and online processing of multiword sequences: An eye-tracking study. *Second Language Research*, *33*, 519–549.
- Yu, L., Cutter, M. G., Yan, G., Bai, X., Fu, Y., Drieghe, D., & Liversedge, S. P. (2016). Word n+2 preview effects in three-character Chinese idioms and phrases. *Language, Cognition and Neuroscience*, *31*, 1130–1149.
- Zhang, B., & Peng, D. (1992). Decomposed storage in the Chinese lexicon. In H. C. Chen & O. J. L. Tzeng. (Eds.), *Language processing in Chinese* (Vol. 90, 1st ed., pp. 131–149). Amsterdam: North-Holland.
- Zhou, J., Ma, G., Li, X., & Taft, M. (2017). The time course of incremental word processing during Chinese reading. *Reading and Writing*. Advance online publication. doi: [10.1007/s11145-017-9800-y](https://doi.org/10.1007/s11145-017-9800-y)

Cite this article: Jiang, S., Jiang, X., and Siyanova-Chanturia, A. (2020). The processing of multiword expressions in children and adults: An eye-tracking study of Chinese. *Applied Psycholinguistics* *41*, 901–931. <https://doi.org/10.1017/S0142716420000296>