Article

second language research

Second Language Research
2017, Vol. 33(4) 519–549

DOI: 10.1177/0267658317708009
journals.sagepub.com/home/slr
SAGE

Frequency, contingency and online processing of multiword sequences: An eye-tracking study

Wei Yi
University of Maryland, USA

Shiyi Lu
Peking University, China

Guojie Ma
Shaanxi Normal University, China

Abstract

Frequency and contingency are two primary statistical factors that drive the acquisition and processing of language. This study explores the role of phrasal frequency and contingency (the co-occurrence probability/statistical association of the constituent words in multiword sequences) during online processing of multiword sequences. Meanwhile, it also examines language users' sensitivity to the two types of statistical information. Using the eye-tracking paradigm, native and advanced nonnative speakers of Chinese were instructed to read 80 disyllabic two-word Chinese adverbial sequences embedded in sentence contexts. Eye movements of the participants were recorded using both early and late measures. Mixed-effects modeling revealed that both phrasal frequency and contingency influenced the processing of the adverbial sequences; however, they were likely to function in different time windows. In addition, both native and nonnative speakers were sensitive to the phrasal frequency and contingency of the sequences, though their degrees of such sensitivity differed. Our findings suggest that adult language learners retain the statistical learning ability in second language acquisition and they may share a general statistical learning mechanism with native speakers when processing multiword sequences.

Keywords

contingency, eye-tracking, multiword sequences, phrasal frequency

Corresponding author:
Shiyi Lu, School of Chinese as a Second Language, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing, 100871, China
Email: lushiyi@pku.edu.cn

I Introduction

1 Usage-based approaches and statistical learning

How people learn a language remains unclear and the underlying language acquisition mechanism has been under debate for a long time (Saffran, 2003). The Universal Grammar (UG) theory raised by Noam Chomsky argues that language is an independent system distinct from other cognitive domains, and that human beings are born with an innate language acquisition device. The UG theory adopts a words-and-rules approach (Pinker, 1999; Pinker and Ullman, 2002) and divides language into several distinct components such as syntax, semantics and phonology. Although such views of language and language acquisition have dominated for over four decades, theoretical alternatives such as those following usage-based approaches have gained more and more favor among language researchers and cognitive scientists in recent years. 'Usage-based approaches' is an umbrella term that incorporates various kinds of theories such as usage-based theories (Bybee, 1998; Ellis, 2003, 2006a, 2006b, 2008; Goldberg, 1995, 2006; Langacker, 1987; Tomasello, 2003), connectionist models (Christiansen and Chater, 1999; Elman, 1990; MacWhinney, 1998; Rumelhart and McClelland, 1986) and exemplar-based models (Bod, 2006; Pierrehumbert, 2001). Distinct from the UG theory, usage-based approaches share the three ideas as follows. First, language and language acquisition are shaped by language use instead of a pre-existing innate system. Second, language is an inventory of symbolic units rather than a combination of words and rules. Such units are called 'constructions', which are form–meaning mappings that relate specific linguistic patterns with certain semantic, pragmatic and discourse functions (Goldberg, 1995, 2006). Third, rules of form–meaning mappings are not explicitly given, but rather emerge from repeated language use based on certain cognitive and psychological mechanisms (Bod, 1998).

Language is abundant in statistical regularities (Ellis, 2006a, 2006b). Thus, learning a language heavily involves figuring out the underlying statistics, and statistical knowledge should also be part of our language knowledge (Gries and Ellis, 2015). This is supported by mounting evidence from the literature of statistical learning. Statistical learning refers to the process by which language users discover the structure of the language input based on its distributional properties (Frost et al., 2015; Rebuschat, 2013) such as frequency, variability, distribution and co-occurrence probability (Erickson and Thiessen, 2015). By keeping track of the underlying distributional information, language users can boost different aspects of language processing, including phonological learning (e.g. Maye et al., 2008; Thiessen and Saffran, 2003), word segmentation (e.g. Saffran et al., 1996; Swingley, 2005), syntactic learning (e.g. Thompson and Newport, 2007; Tomasello, 2001) and category formation (e.g. Gomez and Gerken, 2000). Moreover, statistical learning has been found in both young children (e.g. Gomez and Gerken, 2000; Saffran et al., 1996) and adults (e.g. Frank et al., 2010; Zuhurudeen and Huang, 2016), in both first (e.g. Saffran et al., 1996) and second language acquisition (e.g. Frost et al., 2013; Hamrick, 2014), and in studies using both artificial (Conway et al., 2010; Saffran et al., 1996) and natural languages (Fine and Jaeger, 2013; Zuhurudeen and Huang, 2016). By tracking different types of statistical information, language users can continuously store

and modify representations of constructions at the level of morphemes, words or multiword sequences. From this perspective, language acquisition is essentially an 'intuitive statistical learning problem' (Ellis, 2008: 374).

2 Usage-based approaches and statistical learning of multiword sequences

Usage-based approaches emphasize that constructions – conventionalized form–meaning mappings at varying linguistic levels – are the building blocks of language (Goldberg, 2006). Moreover, constructions are acquired by statistically abstracting the patterns of form–meaning correspondence based on usage events (Ellis et al., 2014). When it comes to multiword sequences, there are good reasons to believe that such larger-than-word units should be represented and processed similarly to other linguistic units, and the acquisition and processing of multiword sequences should also be subject to common statistical learning mechanisms.

Multiword sequences (MWS) – sometimes also referred to as 'formulaic language' (Wray, 2002) – are recurring sequence patterns comprised of multiple words. Generally, MWS are defined based on how frequently a word combination occurs in corpora (Biber et al., 1999) and whether the co-occurrence of the constituent words is random. MWS cover a variety of linguistic phenomena, including idioms (*kick the bucket*), phrasal verbs (*take off*), speech formulae (*what's up?*), irreversible binomials (*bride and groom*), collocations (*make progress*) and lexical bundles (*is one of the*). Although different subcategories of MWS vary in terms of length, idiomaticity and fixedness, current research has shown that they are widely used (Biber et al., 1999; Erman and Warren, 2000), and play a critical role in the development of language fluency (Pawley and Syder, 1983; Wood, 2002) and native-likeness (Pawley and Syder, 1983). In addition, empirical studies also found that highly frequent MWS enjoy certain processing advantages over novel expressions (Arcara et al., 2012; Bannard and Matthews, 2008; Conklin and Schmitt, 2008; Durrant and Doherty, 2010; Ellis et al., 2008; Jiang and Nekrasova, 2007; Siyanova-Chanturia et al., 2011a, 2011b; Tremblay and Baayen, 2010; Tremblay et al., 2011; Underwood et al., 2004).

Two types of statistical information can play important roles in the acquisition and processing of MWS: frequency and contingency (Gries and Ellis, 2015). It is worth mentioning that contingency has been predominantly used in the literature on associative learning (Shanks, 1995), while its use in the field of second language acquisition is a recent development (Ellis, 2006a, 2006b; Ellis et al., 2014; Gries and Ellis, 2015). The human mind is said to be able to implicitly or explicitly acquire the knowledge of statistical correlations between stimulus pairings or the predictive relationships between stimuli and responses, and such processes are called 'contingency learning' (Schmidt, 2012). Language acquisition can be understood as contingency learning (Ellis, 2006a, 2006b) in the sense that language learners must figure out the reliability of form–meaning/function mappings or the strength of the statistical association between linguistic elements (Gries and Ellis, 2015). Using contingency information, language users can get the interpretations that are most relevant to the context and predict what is most likely to be heard or

seen next. In this article, the frequency of MWS was measured by the number of occurrences of the whole word combination in a corpus. On the other hand, the contingency of MWS was operationalized as the probabilistic/predictive relationship between the constituent words in a sequence, which can be measured by a variety of probabilistic measures (Ellis et al., 2014; Gries and Ellis, 2015). Technically, when talking about the statistical association of MWS, 'contingency' can be used interchangeably with other terms such as 'probability' or 'predictability'. However, the term 'contingency' is rooted in the view that language acquisition is associative and statistical, and was treated as a higher-order construct distinguished from its measurement tools (i.e. corpus-based, probabilistic statistics). To maintain the theoretical consistency and avoid conceptual confusion, 'contingency', but not 'probability' or 'predictability', was used in this article. Given that frequency and contingency lie at the core of the statistical learning mechanism for MWS, their role in the processing of MWS – as well as language users' sensitivity to these statistics – are reviewed in the following sections.

3 Frequency and the processing of MWS

Frequency is the most robust statistic among the many kinds of distributional information to which language users are sensitive. Frequency determines how likely a construction is to be experienced by language users, how firmly it is entrenched in the mind, and how readily and automatically it will be accessed and processed (Gries and Ellis, 2015). According to Ellis (2002), language users are intimately tuned to the input frequency, and frequency effects exist in the processing of almost every aspect of language (Diessel, 2007; Ellis, 2002; Jurafsky, 2003). In terms of lexical processing, both comprehension (e.g. Balota et al., 2004; Duyck et al., 2008) and production (e.g. Jescheniak and Levelt, 1994) studies have shown that language users respond faster to high-frequency words than to low-frequency ones. Moreover, such effects exist in both open- and closed-class words (Segui et al., 1982), and among both monolingual and bilingual speakers (Duyck et al., 2008).

Frequency is an indicator of language use. From the perspective of usage-based theories, effects of frequency should exist in linguistic units of varying grain sizes and go beyond the single word level. Indeed, a growing literature on MWS shows that frequency effects do extend to MWS. Overall, such investigations follow two different approaches (Arnon and Snider, 2010), depending on whether the target stimuli are restricted to a certain frequency threshold. The threshold-approach studies aimed to test the hypothesis that highly frequent MWS (i.e. formulaic language) are stored and processed as holistic units (Wray, 2002), which renders them processing advantages over novel expressions. Focusing on MWS in the very high end of the frequency continuum, a massive body of research has been done in recent years. Among such studies, highly frequent MWS were usually extracted from corpora based on preset frequency criteria (Biber et al., 1999), while sequences differing in phrasal frequency – yet matched on other properties – were created as control stimuli. Behavioral performance on the MWS and the control stimuli were analysed in terms of reaction time and/or accuracy rates. Once high-frequency strings were found to be processed faster and/or with higher accuracy rates than controlled novel expressions, researchers would claim the existence of the processing

advantage(s) of MWS, in support of the holistic storage/processing hypothesis. Following such a logic, processing advantages have been found in a variety of highly frequent word combinations, including idioms (Conklin and Schmitt, 2008), language formulae (Jiang and Nekrasova, 2007), collocations (Durrant and Doherty, 2010; Sosa and MacFarlane, 2002), irreversible binominals (Arcara et al., 2012) and lexical bundles (Bannard and Matthews, 2008; Ellis et al., 2008; Tremblay and Baayen, 2010; Tremblay et al., 2011; Underwood et al., 2004).

Studies following the threshold approach built their claim of the holistic processing/storage nature of high-frequency MWS on the observed phrasal frequency effects. However, it remains unclear whether such effects can generalize to less frequent MWS. In other words, findings made by the studies focusing on highly frequent MWS do not lead to the conclusion that frequency effects at the multiword level function in the whole continuum. From a usage-based perspective, MWS exist as a continuum in terms of frequency and other statistical properties; therefore, it is worthwhile to test whether phrasal frequency effects extend to more flexible, less frequent MWS as well. Based on this logic, another line of research expands the exploration of phrasal frequency effects by adopting a continuous approach. Different from the threshold approach, it regards phrasal frequency as a continuum and extracts MWS from a wider frequency range. In a comprehension study by Arnon and Snider (2010), four-word compositional expressions varying across the frequency range were divided into three frequency bins (i.e. high, mid and low bins) using different cutoff values. The authors aimed to test: 1) whether phrasal frequency effects exist in MWS; 2) whether such effects can be observed across the entire frequency range; and, 3) whether using continuous frequency data leads to greater statistical power than treating frequency as a discrete variable. Native speakers of English were asked to judge whether the four-word sequences exist in English or not by responding as fast and as accurately as possible. As reported by the authors, frequent MWS were processed significantly faster than the less frequent control phrases, and such processing advantage appeared in all frequency bins. In addition, it was also found that the use of continuous measures of frequency as predictors did generate more reliable results.

Similar results were found by Janssen and Barber (2012). In their study, native speakers of Spanish were presented with visual stimulus displays depicting two superimposed objects or isolated colored objects. In the former situation, participants were required to name the two objects following the noun–noun format (e.g. *martillo* (‘hammer’) – *rana* (‘frog’)). In the latter, they were required to name the isolated objects in its color following the noun–adjective format (e.g. *anillo* (‘ring’) – *rosa* (‘pink’)). The expressions to be produced varied across the whole continuum of phrasal frequency. By manipulating the phrasal frequency and the frequency of the object nouns, naming latencies of the MWS were found to be affected by phrasal frequency after controlling for the frequency of the constituent words.

Effects of phrasal frequency were also obtained among nonnative speakers. In a study carried out by Wolter and Gyllstad (2013), first language (L1) Swedish English learners were required to judge whether certain collocations exist in English or not. Unknown to the participants, some of the collocations had word-by-word translations in Swedish (congruent collocations), while others (incongruent collocations) did not. Statistical analyses revealed that advanced Swedish learners of English were sensitive to collocational

frequency in that they responded faster to more frequent stimuli. Moreover, such frequency effects were independent of the collocations' congruency status.

4 Contingency and the processing of MWS

Frequency is important for the acquisition and processing of MWS, but it is not the only factor (Ellis, 2008). MWS consist of multiple words co-occurring probabilistically in a sequence; therefore, figuring out such probabilistic co-occurrence pattern is also of great importance. As mentioned previously in this article, such probabilistic relationship can be understood as contingency (Gries and Ellis, 2015) and be measured using various statistical association metrics (Gregory et al., 1999; Gries, 2010; Gries and Ellis, 2015), including forward/backward transitional probability (McDonald and Shillcock., 2003; Tremblay and Baayen, 2010), mutual information (Church et al., 1991; Ellis et al., 2008, Durrant and Doherty, 2010), *t*-score (Wolter and Gyllstad, 2011) and ΔP (Gries and Ellis, 2015). All these measures are computed based on a contingency table. As shown in Table 1, there are four possible combinations of events (*a, b, c, d*) given the cue and/or outcome is either present or absent. Take a two-word combination XY for example: *a* refers to the frequency of the word combination XY, *a*+*c* refers to the frequency of the word Y, *a*+*b* is the frequency of the word X, and *a*+*b*+*c*+*d* refers to the total number of words of the whole corpus.

Association measures such as transitional probability, mutual information (MI), *t*-score and ΔP each has their advantages and disadvantages. For example, MI is known for generating very high association scores for low-frequency MWS such as technical terms or fixed expressions (Ellis et al., 2008; Gries, 2010). In comparison, *t*-score returns high association scores for high-frequency word pairs (Gries, 2010). In terms of directionality, forward/backward transitional probability and ΔP are unidirectional, whereas measures such as *t*-score and MI are bi-directional. Specifically, *t*-score and MI account for the mutual predictability between constituent words in MWS, whereas forward/backward transitional probability and ΔP only address the one-way predictability relationship. In this article, MI was used as the measure of contingency of MWS. The reasons are as follows. First, all the MWS in this study were constructed by twenty-four monosyllabic adverbs that are highly homogenous (they barely differ in terms of frequency or visual complexity). Given no direct evidence supporting the uni-directionality of the predictive relationship between the constituent words, a bi-directional measure should be a better choice in that the potential mutual predictability between the two adverbs can be considered. Second, studies have shown that MI is one of the most robust probabilistic measures to which language users are sensitive (Ellis et al., 2008; Gregory et al., 1999), and it is applicable to various types of MWS, including collocations (Durrant and Doherty, 2010) and lexical bundles (Ellis et al., 2008). Third, MWS used in this study are not technical terms or fixed expressions; therefore, it is likely to circumvent the bias of MI as mentioned above. For example, taking the two-word collocation 'common sense',[1] the MI value can be computed using the formula illustrated below. Specifically, *f(xy)* is the collocational frequency (8.6 times per million), *f(x)* is the frequency of the word 'common,' (177.1 times per million), *f(y)* is the frequency of the noun 'sense,' (3.4 times per million), and *N* is the corpus size (the British National Corpus, 112 million words).

Table 1. A contingency table showing the four possible combinations of events.

	Outcome	No outcome	Total
Cue	*a*	*b*	*a+b*
No cue	*c*	*d*	*c+d*
Totals	*a+c*	*b+d*	*a+b+c+d*

Source. Adapted from Ellis, 2006a.
Note. a, b, c, d represent frequencies of each event.

$$MI = \log_2^{\frac{f(xy)\times N}{f(x)\times f(y)}}$$

Before reviewing the literature on contingency and second language acquisition, three things about mutual information are worth mentioning. First, mutual information is a measure of the strength of the statistical association between constituent words in MWS. The higher the MI value is, the stronger the word combination is statistically associated. Second, there is no minimum threshold value for MI (Simpson-Vlach and Ellis, 2010), as MI values provide only comparative information. Lastly, although contingency as measured by MI or other measures is computed based on frequency counts, it is still distinct from phrasal frequency. In a nutshell, phrasal frequency indicates the likelihood that language learners experience certain MWS. By contrast, contingency illustrates the reliability of the co-occurrence patterns (Gries and Ellis, 2015); that is, how reliably one can expect to see or hear the word X given the word *Y* (or vice versa). From a statistical perspective, the correlation between phrasal frequency and contingency should be rather weak. To illustrate this point, in an ongoing study by the first author of the current study, 8,400 adjective-noun collocations ranging across the frequency continuum were extracted from the British National Corpus; the Pearson correlation between MI scores and collocational frequencies was only 0.09.

Given the importance of contingency in language acquisition and processing, it is worthwhile to investigate whether language users are sensitive to the such information underlying the language input. For native speakers, it has been found that contingency information of syllables (Saffran et al., 1996), phrases (Ellis et al., 2008; Gregory et al., 1999; McDonald and Shillcock, 2003) as well as other linguistic structures is indeed stored in their mind. For example, Saffran et al. (1996) found that 8-month-old infants could use the transitional probability between syllables to discover word boundaries in an artificial language after only two minutes of exposure. Concerning adult language users, Gregory et al. (1999) found that probabilistic information of word combinations also functions in speech production in that highly probable collocations were more often shortened in duration and were more likely to have final /t/ or /d/ sounds deleted. Similar effects of contingency in language processing were also revealed by eye movement patterns in natural reading. In a study by McDonald and Shillcock (2003), adult native English speakers were recruited to read ten excerpts of newspaper articles. It was found that both forward and backward transitional probabilities were predictive of participants' first- fixation durations and gaze durations.

Conversely, few studies have been done to examine second language users' sensitivity to the contingency information underlying the second language (L2) input. Ellis et al. (2008) validated the psychological reality of corpus-extracted academic English formulas (e.g. *the value of the*) in a series of comprehension and production experiments. By manipulating the length (three to five words), phrasal frequency (high, middle, and low) and mutual information (high, middle, and low), multiple regression analyses on reaction time and accuracy rates revealed contrasting result patterns between native and nonnative speakers of English: phrasal frequency effects were found only among nonnative speakers, while effects of mutual information were only found among native speakers. Such findings are quite interesting, yet they are also limited by the small sample size, the lack of the consideration of confounding variables (e.g. constituent word frequency, bigram/trigram frequency) and the inadequate automaticity of the experimental tasks in the research design.

In another study by Ellis and his colleagues (Ellis et al., 2014), the processing of English verb–argument constructions (VAC) was examined in order to explore the role of frequency, contingency and semantic prototypicality. Two free association tasks were employed, in which native English speakers were required to generate the first word that come into their mind (Experiment I), or to generate as many verbs as possible in one minute (Experiment II) to fill in the verb slot in 40 VAC frames (e.g. *he* ____ *across the* ...). VAC-verb contingency was measured using ΔP (Gries and Ellis, 2015). For both experiments, frequencies of verb types generated for each VAC were regressed on verb frequency in the VAC, VAC-verb contingency and verb prototypicality in terms of centrality within the VAC semantic network. All these factors were found to be significant, thus confirming the role of contingency as part of the processing mechanism of MWS among native speakers.

Native speakers' sensitivity to the contingency information of MWS is also supported by further evidence. A study by Tremblay and Baayen (2010) explored the holistic representation of MWS by examining the effects of frequency and contingency. The researchers extracted 432 regular four-word sequences (e.g. *becoming increasingly clear that*) ranging from 0.03 to 105 times per million in whole-string frequency from the British National Corpus. Native speakers of English were asked to recall as many of the four-word sequences as possible that were learned in practice sessions without delay. Similar to the previous findings (Ellis et al., 2008), no effects of phrasal frequency were obtained. However, the whole-string contingency (measured by LogitABCD[2]) were found to be predictive of the immediate free recall performance.

5 Summary

Current literature on the acquisition and processing of MWS is limited in a couple of ways. First, only a small number of studies (e.g. Ellis et al., 2008; Ellis et al., 2014; Tremblay and Baayen, 2010) have included both frequency and contingency in their design. Therefore, it is not clear: 1) whether phrasal frequency and contingency impact the processing of MWS in different ways; 2) whether either effect can be observed when controlling for the other; and 3) whether there is any interaction between the two factors. Second, it remains unknown whether language users (especially nonnative speakers) are

sensitive to the phrasal frequency and/or contingency of MWS, and whether such statistical sensitivity differs between native and nonnative speakers. Third, most current studies narrowly focus on extremely high-frequency MWS, leaving it unclear whether MWS across the range of phrasal frequency and/or contingency share the same statistical learning mechanisms. Finally, obtained findings about the processing of MWS may disappear if one switches to more automatic experimental tasks (Durrant and Doherty, 2010). Therefore, studies using more automatic online experimental techniques such as masked priming or eye tracking are needed to verify previous experimental findings.

To get a better understanding about the role of phrasal frequency and contingency during the processing of MWS, as well as native and nonnative speakers' statistical sensitivity to these two kinds of distributional information, the following four steps were taken. First, both phrasal frequency and contingency were incorporated as variables of interest. Second, both native and nonnative speakers were recruited so that their sensitivity to phrasal frequency and contingency can be compared. Third, MWS spreading across the range of phrasal frequency were extracted from the corpus. Fourth, the eye-tracking paradigm was employed and statistical analyses were carried out using data obtained from various kinds of eye movement measures.

II Chinese adverbial sequences

Chinese adverbial sequences were used as stimuli in this study. Chinese is an isolating language that is poor in inflectional morphology and syntactic rules (Portin et al., 2008). Instead, grammatical functions are realized primarily through word order and function words such as adverbs. Chinese adverbs are usually short in length and highly frequent. They can appear in different positions of the sentence and modify verbs, adjectives or the whole sentence. Disyllabic Chinese adverbial sequences consist of two monosyllabic adverbs, yet they can only be placed in the middle of the sentence. To better understand Chinese adverbial sequences, '仍没 (réng méi)' is given as an example. As can be seen, this adverbial sequence consists of two monosyllabic adverbs and is placed before the verb phrase, acting as a modifier. The meaning of the whole sequence is the combination of the meanings of its constituent words (i.e. '仍' and '没'). Functionally, it expresses the present perfect tense as in English; semantically, it describes an uncompleted action that is expected by the speaker.

毕业 将近 半年 了，我 仍 没 找到 工作。[3]

Bìyè jiāngjìn bànnián le, wǒ réng méi zhǎodào gōngzuò.

Graduate-will almost-half a year-'le' (completed action marker), I-still-have not-find-job.

Having graduated almost half a year ago, I still have not found a job.

The adverbial sequences used in this study were comprised of two highly frequent monosyllabic adverbs, and each adverb was represented by a single Chinese character. Corpus linguistic studies have revealed that Chinese adverbial sequences consisting of two monosyllabic adverbs are commonly used (Fang, 2012; Li, 2010). Given the properties of Chinese adverbs and adverbial sequences, there may be several potential

concerns about the appropriateness of their use in this study. For example, one may doubt whether function words are subject to the same effects of frequency and contingency that exists in content words. However, evidence has shown that not only that function words share a common lexical processing mechanism with content words, but also that frequency (e.g. Schmauder et al., 2000) and contingency (Jurafsky et al., 2001; McDonald and Shillcock, 2003) do function in the processing of function words. Another concern may be raised regarding the generalizability of the results found in Chinese adverbial sequences to other types of MWS. The adverbial sequences used in the current study can be categorized as lexical bundles (Ellis et al., 2008; Tremblay and Baayen, 2010). Given that different subcategories of MWS should be subject to common statistical processing mechanisms from the usage-based perspective, we suggest that results obtained from Chinese adverbial sequences can also be generalized to other types of larger-than-word units.

III The eye-tracking paradigm

Eye tracking was used in this study due to the following three considerations. First, metalinguistic knowledge or strategy is less likely to be involved (Rayner, 1998, 2009) when participants are instructed to read for meaning in an eye-tracking experiment. Second, the eye-tracking technique provides extremely rich data collected from a variety of eye movement measures. Most importantly, by incorporating different temporal eye movement measures, both early and late stages of processing can be revealed (Roberts and Siyanova-Chanturia, 2013). Early measures, including first fixation duration and first pass reading time, are indicative of early processes during reading, such as familiarity checks, access to orthographic/phonological information and lexical meaning (Reichle et al., 1998; Roberts and Siyanova-Chanturia, 2013). Comparatively, late measures such as total reading time, second-pass reading time and fixation count are believed to reflect later processes, such as reanalysis of information, integration of information in discourse and recovery from processing difficulties (Paterson et al., 1999; Rayner et al., 1989). Third, eye movement measures also have been found to be very sensitive to frequency as well as contingency (Engbert et al., 2005; McDonald and Shillcock, 2003; Reichle et al., 1998).

When designing an eye-tracking study, the most important consideration is to choose the appropriate measures. As Rayner (1998) argued, it is never a good practice to use only one single measure, since '…any single measure of processing time per word is a pale reflection of the reality of cognitive processing' (Rayner, 1998: 377). For the purpose of this study, five measures – including both early (first fixation duration, first pass reading time) and late (total reading time, fixation count, skipping rate) measures – were used. Rayner (1998) also suggested that when the unit of analysis is larger than a word, distinctions should be made between first-pass and second-pass reading time. However, we decided not to include second-pass reading because the adverbial sequences used in this study were about the average length of Chinese words; consequently, we expect that their overall reading pattern should be similar to single words. This reasoning also explains why we adopted first fixation duration in our analyses. Roberts and Siyanova-Chanturia (2013) claimed that first fixation duration is useful only when the region of interest is a word; for larger-than-word linguistic units, it is not suitable because the

probability of further fixations will increase. Given that the adverbial sequences are of the average length of Chinese words, we believe that it is worthwhile to keep first fixation duration as a measure.

Unlike alphabetical languages such as English, Chinese is logographic in terms of the writing system. In addition, Chinese words are not spatially segmented, and they are represented by characters that differ in visual complexity (the number of strokes). In spite of these differences, the fundamental nature of the reading of Chinese is similar to that of alphabetic languages (Li et al., 2014; Zang et al., 2011). For example, like English, Chinese reading is basically word-based (Li et al., 2014; Zang et al., 2011). In previous eye-tracking research, the visual complexity of Chinese characters was found to influence the fixation duration (e.g. Ma and Li, 2015), yet such an effect is modulated by word frequency (Liversedge et al., 2014). In addition, similar to English reading, Chinese readers are also able to exploit the predictability information when reading (Rayner et al., 2005).

IV The present study

To examine the role of phrasal frequency and contingency during the online processing of MWS, as well as the sensitivity to such statistical information among native and nonnative speakers, the following research questions are addressed:

1. Controlling for the effect of contingency, is there any effect of phrasal frequency during the online processing of Chinese adverbial sequences among native and/or nonnative speakers?
2. Controlling for the effect of phrasal frequency, is there any effect of contingency (measured by MI) during the online processing of Chinese adverbial sequences among native and/or nonnative speakers?
3. Is there any interaction effect between phrasal frequency and contingency during the online processing of Chinese adverbial sequences among native and/or nonnative speakers?
4. Are native and nonnative speakers of Chinese sensitive to different statistical information (i.e. phrasal frequency and contingency) when processing Chinese adverbial sequences?

V Method

1 Participants

Twenty native Chinese speakers (8 males, 12 females) and twenty nonnative Chinese speakers (8 males, 12 females) participated in this study. All participants were college students recruited from universities in Beijing, China. Nonnative speakers of Chinese came from a wide range of L1 background, including Arabic (2), Dutch (1), English (6), German (1), Kazakh (1), Persian (1), Russian (4), Spanish (3). Half of the nonnative speakers were master students majoring in TCSOL (Teaching Chinese to Speakers of Other Languages) programs, and the other half were students taking Chinese courses at

Table 2. Nonnative speakers' Chinese learning background (n = 20).

	M	Minimum	Maximum	SD
Age (years)	23.9	20	34	3.8
Duration of formal Chinese instruction (months)	52.5	32	96	19.7
Self-rating				
Listening	7.8	7.0	9.0	0.8
Reading	7.3	6.0	9.0	1.3
Speaking	7.5	6.0	9.0	0.8
Writing	6.7	5.0	8.0	1.2

advanced levels. By the time of the experiment, all nonnative speakers had passed the second-highest level of the Chinese Proficiency Test (HSK-5), which is comparable to the CEFR level C1. Regarding the age of onset, none of them started learning Chinese before the age 15. The duration of their formal Chinese instruction varied from 32 to 96 months, and the average duration of their Chinese learning was 52.5 months. L2 learners' self-ratings of their Chinese language proficiency based on a 10-point scale were relatively high (Table 2). All participants had normal or corrected to normal vision.

2 Materials and design

The stimuli were created in the following way. First, 33 most frequent monosyllabic Chinese adverbs were selected after consulting the *Classified word frequency list of Modern Chinese* (National Committee of Language and Script, 2015a). Then familiarity ratings of the adverbs were collected from five advanced learners of Chinese based on a four-point scale. Nine monosyllabic adverbs that received an average rating of less than two were excluded. Subsequently, all possible two-word sequences constructed by the remaining 24 monosyllabic adverbs (Appendix 1) were listed and searched in the CCL Modern Chinese Corpus (Center for Chinese Linguistics, Peking University, 2015). Adverbial sequences that occurred at least one time per million words in the CCL corpus were selected, resulting in 119 candidates.

Chinese has no explicit word boundaries. Therefore, sequences of Chinese characters need to be segmented into words using a tokenizer (Gries and Ellis, 2015). However, word-segmentation is not common for large-scale Chinese corpora due to the enormity of the task. The CCL corpus consists of 581 million Chinese characters and is not segmented. Given such a situation, the corpus size of the CCL corpus has to be estimated. Since few studies on the character-to-word ratio of Chinese can be referred to, this estimation was carried out based on the character-to-word ratios of two other large-scale, segmented Chinese corpora: the Modern Chinese Corpus (National Committee of Language and Script, 2015b) and the Balanced Corpus of Modern Chinese (Academia Sinica, 2015). The Modern Chinese Corpus consists of 19,455,328 characters that were segmented into 12,842,116 words (character-to-word ratio: 1.515:1), while the Balanced Corpus of Modern Chinese consists of 7,949,851 characters that were segmented into

Table 3. A summary of the characteristics of the experimental stimuli.

Characteristics	Condition			
	HF–HMI	HF–LMI	LF–HMI	LF–LMI
	将-不 jiang-bu will-do not	也-还 ye-hai also-again	仍-很 reng-hen still-very	真-没 zhen-mei really-not yet
Phrasal frequency	20.3 (11.2)	19.3 (17.1)	3.5 (2.4)	3.8 (1.8)
MI	4.8 (1.7)	2.3 (0.7)	4.0 (1.0)	2.1 (0.7)
Initial word frequency	1,483.7 (1,174.3)	1,882.4 (1521.9)	858.3 (768.8)	1,320.7 (875.6)
Terminal word frequency	1,569.3 (1,830.9)	3,127.1 (1,976.1)	643.2 (584.6)	1,349.0 (1,367.5)
Initial word strokes	6.0 (3.3)	6.6 (3.5)	7.8 (2.3)	7.3 (3.6)
Terminal word strokes	6.9 (2.9)	5.4 (2.6)	7.4 (2.6)	7.4 (2.9)
Total strokes	12.9 (3.8)	11.9 (3.9)	15.1 (3.7)	14.7 (4.4)
Sequence familiarity	2.9 (0.4)	3.3 (0.4)	3.2 (0.4)	2.8 (0.4)

Note. Frequency was measured by the number of occurrences per million words. Each example in the table first presents the Chinese adverbial sequence in a word-by-word fashion, then the Chinese pinyin (e.g. *jiang-bu*) and the English translation (e.g. *will-not*). Standard deviations are given in the parentheses following the means. HF–HMI: high frequency – high mutual information; HF–LMI: high frequency – low mutual information; LF–HMI: low frequency – high mutual information; LF–LMI: low frequency – low mutual information.

4,892,324 words (character-to-word ratio: 1.625:1). The average character-to-word ratio is 1.57:1. Using this ratio, the CCL corpus was estimated to be 300 million words after excluding all non-Chinese characters.

Familiarity ratings of the 119 candidate sequences were also collected based on a four-point scale (1 → 4: 'I've never seen this before' → 'I am very familiar with this'), such that sequences receiving an average familiarity rating of less than 2 were removed. Based on the frequency data obtained from the CCL corpus, MI values for all adverbial sequences were computed.[4] Using a stratified sampling method, 80 MWS were then chosen to represent the two levels on both phrasal frequency (high vs. low, cutoff point[5] at 7 times per million) and MI (high vs. low, cutoff points at 3.0). Sequences were grouped into four conditions, and their characteristics are shown in Table 3.

The 80 adverbial sequences were then paired in groups of four so that each group consists of sequences from the four conditions (high frequency – high MI, high frequency – low MI, low frequency – high MI, low frequency – low MI). Each group of the sequences were then embedded in the four different sentence frames, generating four different stimuli lists. For each sentence frame, the content before the adverbial sequence was the same. Moreover, the word closely following the adverbial sequence was matched in terms of frequency across the four sentences under the same sentence frame. When writing the sentences, the following four principles were followed. First, the target sequences were embedded neither in the initial nor the terminal portion of the sentences, and it was ensured that there were at least six Chinese characters before

Table 4. Examples of sentences used in the study.

Condition	Example sentence
HF–HMI	直到十九世纪初人们仍不清楚地球上有多少生物。 Until the beginning of the 19th century, people still did not know the total number of the life on the earth.
HF–LMI	直到十九世纪初人们才不怀疑这项新的科学技术。 Until the beginning of the 19th century, people only did not doubt this new technology.
LF–HMI	直到十九世纪初人们仍很害怕火车这种交通工具。 Until the beginning of the 19th century, people still very much feared the train as a transport.
LF–LMI	直到十九世纪初人们仍没解决食品安全的问题。 Until the beginning of the 19th century, people still had not solved the problem of food safety.

Note. The underlined Chinese characters indicate the adverbial sequences, and the underlined English words are their literal translations.

and after the target sequences. Second, to maintain the readability, all sentences were written using words or characters below the level of HSK-5. Third, to make sure that sentence contexts would not impede or promote the understanding, sentences were written in neutral contexts, and the neutrality of the sentences were ensured by three graduate students from Peking University. Finally, the length of the experimental sentences was strictly controlled, ranging from 20 to 22 Chinese characters ($M = 20$, $SD = 0.6$). Examples of the experimental sentences are shown in Table 4. In total, 320 target sentences along with 80 filler sentences were created. Participants were randomly assigned to read one of the four lists of critical sentences. In addition, 6 practice trials were made, leading to a total of 166 sentences (i.e. 80 critical sentences, 80 filler sentences, 6 practice sentences) to be read for each participant. To make sure that participants read the sentences attentively, nearly one-third of the sentences were followed by comprehension questions, and responses to them were required by choosing the best answer out of three choices.

3 Apparatus and procedure

Sentences were presented in a normal, unspaced manner on a 21-inch CRT monitor (resolution: 1,024 × 768 pixels; refresh rate: 150 Hz) that was connected to a Dell PC. Each sentence was displayed in a single line with Song 20-point font, and the characters were shown in black (RGB: 0, 0, 0) on a gray background (RGB: 128, 128, 128). Participants seated at a viewing distance of 580 mm from the computer monitor, with their head stabilized by means of a chin rest and a forehead rest. At the viewing distance, each character subtended a visual angle of approximately 0.7°. Participants read sentences binocularly, but only the right eye was monitored. Eye movements were recorded using the Eyelink 1000 system, and the sampling rate was 1,000 Hz.

Participants were tested individually. Written instructions about the experiment were given to them after entering the lab. To ensure that participants feel comfortable and that the eye-tracker captures the eye-movement, the height of the chin rest and/or the chair were adjusted when needed. Before starting the experiment, the eye-tracker was calibrated using a three-point calibration, associated with a validation procedure. The maximal error of the validation was 0.5 degrees in visual angle. The experiment can be paused whenever the participants feel the need to have a rest. In the case of resuming the experiment, another set of calibration and validation was carried out following the same procedure as mentioned above. Participants were required to read the sentences silently for meaning at their own pace. After the participant successfully fixated on a character-sized box at the location of the first character of the sentences, a sentence would be presented. Once finishing reading each sentence, a button was pressed to proceed. Participants responded to the comprehension questions by pressing the button for the correct answer. The experiment took about 30 minutes for the native Chinese speakers and 45 minutes for the nonnative speakers.

VI Statistical analyses

Analyses were conducted using mixed-effects models with crossed random effects for subjects and items using R (version 3.3.1; R core team, 2016). Specifically, linear mixed effects models were built to analyse the first fixation duration (FFD), first-pass reading time (FPR) and total reading time (TRT) data, whereas mixed-effect Poisson models and mixed-effects logistic models were fit to analyse the fixation count (FXC) and fixation probability (FXP) data respectively, using the *lme4* package (version 1.1-12, Bates et al., 2015). Independent variables of interest included phrasal frequency and mutual information of the adverbial sequences and speaker (natives vs. nonnatives of Chinese). Word-level properties including the frequency of the two constituent words (word1frequency, word2 frequency) and the visual complexity of the Chinese characters representing the two constituent words (the number of strokes of word1/word2) were treated as covariates. All variables were continuous except speaker. Frequencies (phrasal frequency, word1/word2 frequency) and reaction times (FFD, FPR, TRT) were logged (natural log). Medium correlations were found between word2 length and word2 frequency ($r = -.56$, $p < .05$), as well as between word2 frequency and MI ($r = -.56$, $p < .05$). To reduce the problem of collinearity, all continuous predictors (i.e. phrasal frequency, MI, word1 frequency, word2 frequency, word1 strokes, word2 strokes) were centered at their means. The categorical variable 'speaker' was dummy-coded using native speakers as the reference level.

Models were fit using a maximum likelihood technique and they were built based on the following procedure. At first, a preliminary model[6] was built in which Phrasalfreq (log-transformed and centered phrasal frequency), MI (centered mutual information) and Speaker (dummy-coded) were entered as fixed effects. As argued by Barr et al. (2013), confirmatory studies aiming to test theoretically based hypotheses should always keep a maximal random effects structure justified by the design, so that they have model results that can generalize best. Following this idea, random effects were fit using a maximal

random effects structure that included random intercepts for subjects and items, by-subject random slopes for Phrasalfreq, MI and their interactions, as well as by-item random slope for Speaker. By-subject random slopes allow for differences among subjects in terms of their degree of sensitivity to phrasal frequency, MI and their interaction, whereas by-item random slope allows each experimental stimulus (item) to function differently depending on whether the participant is a native or nonnative speaker. Later, a series of four covariate models were built by gradually adding each of the covariates (i.e. Word1freq/Word2freq: log-transformed and centered constituent word frequencies; Word1/Word2strokes: centered constituent word strokes). Model comparisons were made between the preliminary model and the covariate models following a forward model selection procedure by using the *anova()* function in the *lme4* package. *p* values for the linear mixed-effects models were estimated based on the *t* distribution using the formula[7] raised by Baayen (2008). Results of the final best-fitting model for each dataframe (FFD, FPR, TRT, FXC, FXP) were reported in the following section (*alpha*-levels were set at .05).

VII Results

Prior to the analyses, all trials in which there were one or more blinks within the region of interest (i.e. the adverbial sequences) or in which there were three or more blinks in the whole sentence were excluded. This led to a loss of 3.4% of the data. The overall accuracy rates of native and nonnative speakers for the comprehension questions were 96% and 94% respectively, suggesting that participants had no difficulty answering the comprehension questions. One native speaker was removed from analysis because of his extremely low accuracy rate (33%). Eye movements were analysed based on data collected from five measures: first fixation duration (FFD), first pass reading time (FPR), total reading time (TRT), fixation count (FXC), and fixation probability (FXP).

1 First fixation duration results

Following the procedure as described in the previous section, a preliminary linear mixed-effects model was constructed and then compared with the covariate models. Model comparisons showed that the preliminary model fit best. Results are reported in Table 5. The effect of Speaker was found to be significant (estimate = 0.16, $SE = 0.06$, $t = 2.79$, $p = .005$), indicating that nonnative speakers of Chinese read the sequences significantly slower than natives. The mean reading times for the two groups can be computed using the exponential function: $M_{\text{natives}} = \exp(5.49) = 242.3$ ms; $M_{\text{nonnatives}} = \exp(5.49 + 0.16) = 284.3$ ms. Additionally, the effect of Phrasalfreq was marginally significant (estimate = −0.02, $SE = 0.01$, $t = -1.80$, $p = .072$). This suggests that higher-frequency sequences were read faster by both groups of participants in terms of first-fixation duration.

2 First-pass reading time results

Model comparisons between the preliminary and the covariate linear mixed-effects models for FPR (logged) suggested that the covariate model that included Word1freq,

Table 5. Linear mixed-effects model results for first fixation duration (in logged milliseconds).

Parameters	Fixed effects			Random effects			
				By subject		By item	
	Estimate	*SE*	*t*	Variance	*SD*	Variance	*SD*
Intercept	5.49	0.04	135.58***	0.03	0.17	0.00	0.05
Phrasalfreq	–0.02	0.01	–1.80	0.00	0.01	–	–
MI (mutual information)	0.01	0.01	1.58	–	–	–	–
Speaker	0.16	0.06	2.79*	–	–	0.00	0.05
Phrasalfreq × MI	–0.01	0.01	–0.79	–	–	–	–
Phrasalfreq × Speaker	0.01	0.01	0.43	–	–	–	–
MI × Speaker	–0.01	0.01	–0.67	0.00	0.01	–	–
Phrasalfreq × MI × Speaker	0.00	0.01	–0.01	–	–	–	–

Note. There are 3,000 observations, where one observation is equal to one RT measurement for one adverbial sequence read by one participant. Model formula: FFD (logged) ~ Phrasalfreq*MI*Speaker + (1 + Phrasalfreq*MI| Subject) + (Speaker | Item). * $p < .05$; ** $p < .01$; *** $p < .001$.

Word1strokes and Word2strokes fit best. Results are summarized in Table 6. An abundance of effects were significant, including the effects of Phrasalfreq (estimate = 0.09, $SE = -0.04$, $t = -2.18$, $p = .029$), Speaker (estimate = 0.43, $SE = -0.07$, $t = 6.09$, $p < .001$), two-way interactions between Speaker and Phrasalfreq (estimate = −0.38, $SE = 0.04$, $t = -8.90$, $p < .001$), between Speaker and MI (estimate = −0.13, $SE = 0.02$, $t = -6.68$, $p < .001$), and a three-way interaction between Phraslfreq, MI and Speaker (estimate = 0.04, $SE = 0.02$, $t = 2.32$, $p = .020$). Furthermore, two word-level effects were also found significant: Word1freq (estimate = −0.07, $SE = 0.01$, $t = -6.58$, $p < .001$) and Word1strokes (estimate = 0.01, $SE < 0.001$, $t = 2.66$, $p = .007$). These word-level effects support that adverbial sequences with higher-frequency initial words were read faster, and those with visually more complex (i.e. more strokes) initial words took longer time to process. The average FPR for native and nonnative speakers can be computed using the exponential function: $M_{\text{natives}} = \exp(5.52) = 249.6$ ms; $M_{\text{nonnatives}} = \exp(5.52 + 0.43) = 383.8$ ms.

Given that a three-way interaction effect between Phrasalfreq, MI and Speaker was found, the sequence-level (Phrasalfreq) and the subject-level (Speaker) fixed effects as well as the two-way interactions (Phrasalfreq × MI, MI × Speaker) must be examined based on the overall interaction pattern (Phrasalfreq × MI × Speaker). The three-way interaction effect was plotted using the *effects* package (version 3.1-2; Fox, 2003) as shown in Figure 1. Since Phrasalfreq and MI are continuous and Speaker is categorical, the three-way interaction effect plot was split into two sets based on each level of Speaker (natives vs. nonnatives). To illustrate the interaction between Phrasalfreq and MI, a group of four graphs was plotted for each type of speakers, dividing the continuous variable MI (centered) into four intervals. The medians of the MI intervals were –2, 0, 2 and 4. For each of the graphs, the horizontal axis was Phrasalfreq (logged and centered), the vertical axis was FPR (logged) and the predicted regression line was given.

Table 6. Linear mixed-effects model results for first-pass reading time (in logged milliseconds).

Parameters	Fixed effects			Random effects			
				By subject		By item	
	Estimate	*SE*	*t*	Variance	*SD*	Variance	*SD*
Intercept	5.52	0.06	99.26***	0.03	0.18	0.01	0.07
Phrasalfreq	0.09	0.04	2.18*	0.00	0.05	–	–
MI (mutual information)	0.00	0.02	–0.02	0.00	0.03	–	–
Speaker	0.43	0.07	6.09***	–	–	0.01	0.11
Word1freq	–0.07	0.01	6.58***	–	–	–	–
Word1strokes	0.01	0.00	2.66**	–	–	–	–
Word2strokes	0.02	0.00	6.50***	–	–	–	–
Phrasalfreq × MI	–0.02	0.02	–1.17	0.00	0.00	–	–
Phrasalfreq × Speaker	–0.38	0.04	8.90***	–	–	–	–
MI × Speaker	–0.13	0.02	6.68***	–	–	–	–
Phrasalfreq × MI × Speaker	0.04	0.02	2.32*	–	–	–	–

Note. There are 2,989 observations, where one observation is equal to one RT measurement for one adverbial sequence read by one participant. Model formula: FPR (logged) ~ Phrasalfreq*MI*Speaker + Word1freq + Word1strokes + Word2strokes + (1 + Phrasalfreq*MI | Subject) + (Speaker | Item). * $p < .05$; ** $p < .01$; *** $p < .001$.

As shown in the plot, the slopes of the regression lines varied in terms of positivity and steepness. The slopes illustrated the effects of Phrasalfreq on logged FPR. For native speakers, the slope at each MI interval was positive, meaning that the more frequent the adverbial sequences were, the more time-consuming they were processed in terms of FPR. This finding is quite puzzling, because higher-frequency MWS generally should be processed faster than lower-frequency ones. On the other hand, when visually examined, the steepness of the slopes of Phrasalfreq seems to decrease as MI values grows, indicating that the effects of phrasal frequency might have been attenuated for sequences of higher MI values. However, model fit results of the three-way interaction effect suggested that such attenuation was not significant, as no significant interaction effect between phrasal frequency and mutual information existed for native speakers (estimate = −0.02, $SE = 0.02$, $t = -1.17$, $p = .242$). By contrast, for nonnative speakers, the slope at each MI interval was negative, suggesting that more frequent adverbial sequences were read faster. In addition, the steepness of the slopes of Phrasalfreq in Figure 1 also seemed to decrease as MI values increase, indicating that the facilitative effect of phrasal frequency might have been attenuated for sequences of higher MI values. This was confirmed by the significant three-way interaction effect between Phrasalfreq, MI and Speaker (estimate = 0.04, $SE = 0.02$, $t = 2.32$, $p = .020$). Moreover, given that the MI effect was not significant (estimate = −0.0003, $SE = 0.02$, $t = -0.02$, $p = .984$), the significant two-way interaction effect between MI and speaker (estimate = −0.13, $SE = 0.02$, $t = -6.68$, $p < .001$) suggested that only nonnative speakers of Chinese were sensitive to the mutual information during the first-pass reading of the adverbial sequences. To be

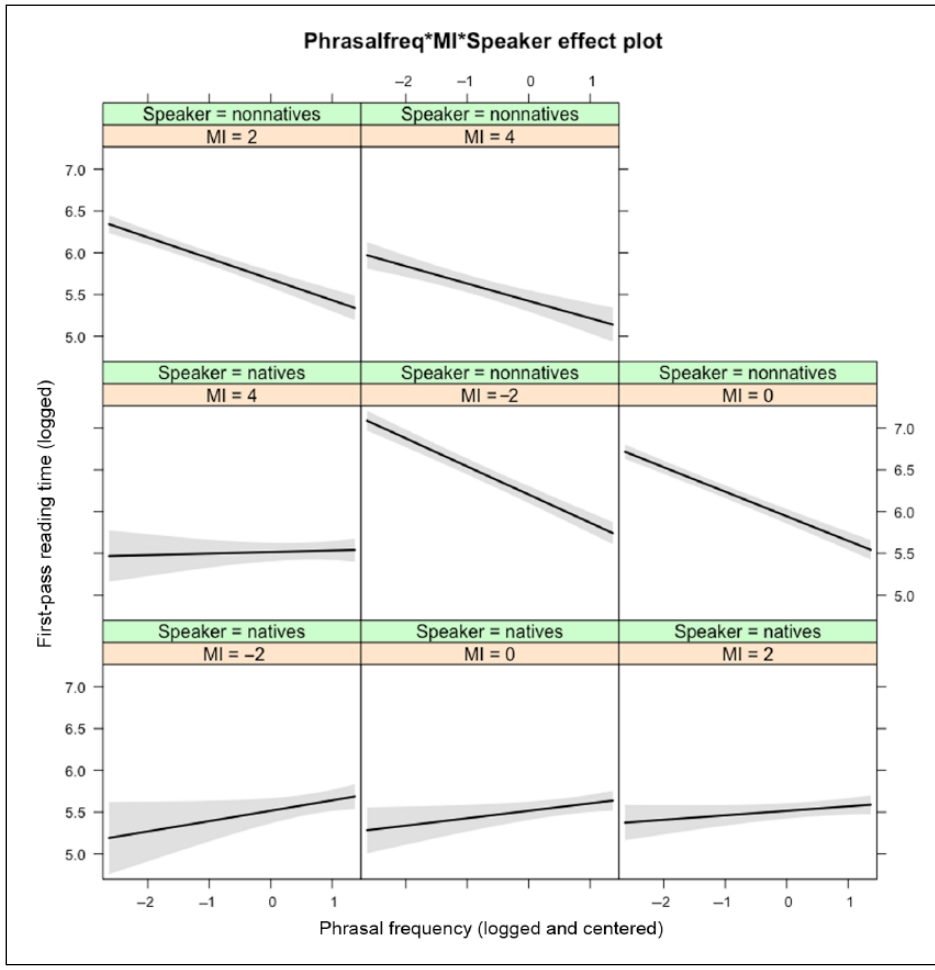

Figure 1. Three-way interaction effect between Phrasal Frequency, MI and Speaker on FRP.

specific, the negative estimate (–0.13) suggested that for each unit increase of centered MI, the first-pass reading time for the sequences would decrease by about 13% [1 – exp(–0.13)].

3 *Total reading time results*

Model comparisons following the described procedure found that the preliminary model fit best for TRT. Overall, the result pattern (Table 7) was similar to that in FFD. Native speakers of Chinese read the adverbial sequences significantly faster than nonnative speakers (estimate = 0.61, $SE = 0.10$, $t = 5.89$, $p < .001$). The average TRT (in milliseconds scale) for the two groups can be computed using the exponential function: $M_{\text{natives}} =$

Table 7. Linear mixed-effects model results for total reading time (in logged milliseconds).

Parameters	Fixed effects			Random effects			
				By subject		By item	
	Estimate	*SE*	*t*	Variance	*SD*	Variance	*SD*
Intercept	5.80	0.07	78.32***	0.10	0.31	0.01	0.07
Phrasalfreq	–0.03	0.02	–1.85	0.00	0.03	–	–
MI (mutual information)	0.00	0.01	–0.05	0.00	0.02	–	–
Speaker	0.61	0.10	5.89***	–	–	0.02	0.15
Phrasalfreq × MI	–0.01	0.01	–0.82	–	–	–	–
Phrasalfreq × Speaker	0.00	0.02	–0.21	–	–	–	–
MI × Speaker	0.01	0.01	0.62	0.00	0.01	–	–
Phrasalfreq × MI × Speaker	–0.01	0.01	–0.54	–	–	–	–

Note. There are 3,134 observations, where one observation is equal to one RT measurement for one adverbial sequence read by one participant. Model formula: TRT (logged) ~ Phrasalfreq*MI*Speaker + (1 + Phrasalfreq*MI| Subject) + (Speaker | Item). * $p < .05$; ** $p < .01$; *** $p < .001$.

exp(5.80) = 330.3 ms; $M_{\text{nonnatives}}$ = exp(5.80 + 0.61) = 607.9 ms. The effect of Phrasalfreq was marginally significant (estimate = −0.03, *SE* = 0.02, *t* = −1.85, *p* = .064), indicating that higher-frequency sequences were read faster by both groups of participants.

4 Fixation count results

Mixed-effects Poisson regression models were built to analyse the fixation count data, following the same procedure as mentioned before. Model comparisons revealed that the covariate model of which Word1freq and Word2freq were included fit best. Results are summarized in Table 8. As illustrated, effects of Phrasalfreq (estimate = −9.77, *SE* = 4.82, *z* = −2.03, *p* = .042) and MI (estimate = 6.98, *SE* = 3.34, *z* = 2.09, *p* = .037), as well as the two-way interactions between Phrasalfreq and Speaker (estimate = −0.13, *SE* = 0.06, *z* = −1.97, *p* = .049), between MI and Speaker (estimate = −0.13, *SE* = 0.03, *z* = −3.92, p < .001) were significant. This suggested that the number of fixations made on the adverbial sequences was influenced by phrasal frequency and mutual information independently, regardless of the type of speaker. Specifically, one unit increase in Phrasalfreq (logged and centered) was predicted to lead to a decrease of 9.77 fixation counts (in log scale) on the sequences, while one unit increase in MI (centered) was predicted to result in an increase of 6.98 fixation counts (in log scale). Moreover, the significant two-way interactions suggested that nonnative speakers of Chinese were more sensitive to both the phrasal frequency and MI of the adverbial sequences, as indicated by the estimates (–0.13 for Phrasalfreq × Speaker and MI × Speaker). Finally, the effects of Word1freq and Word2freq were also found to be significant, indicating that sequences with higher-frequency constituent words received more fixation counts.

Table 8. Mixed-effects Poisson model results for fixation count.

Parameters	Fixed effects			Random effects			
				By subject		By item	
	Estimate	*SE*	*z*	Variance	*SD*	Variance	*SD*
Intercept	–0.10	0.11	–0.89	0.16	0.40	0.00	0.06
Phrasalfreq	–9.77	4.82	–2.03*	0.00	0.03	–	–
MI (mutual information)	6.98	3.34	2.09*	0.00	0.04	–	–
Speaker	0.86	0.14	5.97***	–	–	0.02	0.12
Word1freq	9.86	4.82	2.05*	–	–	–	–
Word2freq	9.91	4.82	2.06*	–	–	–	–
Phrasalfreq × MI	–0.02	0.02	–0.71	0.00	0.02	–	–
Phrasalfreq × Speaker	–0.13	0.06	–1.97*	–	–	–	–
MI × Speaker	–0.13	0.03	–3.92***	–	–	–	–
Phrasalfreq × MI × Speaker	0.01	0.03	0.39	–	–	–	–

Note. There are 3,574 observations, where one observation is equal to one FXC measurement for one adverbial sequence read by one participant. Model formula: FXC~ Phrasalfreq*MI*Speaker + Word1freq + Word2freq + (1 + Phrasalfreq*MI | Subject) + (Speaker | Item). * $p < .05$; ** $p < .01$; *** $p < .001$.

5 Fixation probability results

Mixed-effects logistic regression models were built to analyse the fixation probability data, following the same procedure as described before. Model comparisons revealed that the preliminary model that included no covariates fit best. Results are summarized in Table 9. Speaker effect was found to be significant (estimate = 3.07, $SE = 0.59$, $z = 5.20$, $p < .001$), meaning that nonnative speakers of Chinese were far more likely to fixate on the adverbial sequences than native speakers did.

VIII Discussion

This study investigated the role of phrasal frequency and contingency as well as language users' sensitivity to these two statistical properties during the online processing of Chinese adverbial sequences. In answering the research questions, five main findings were revealed. First, after controlling for the effects of word-level factors (Word1freq, Word2freq, Word1strokes, Word2strokes) and contingency, phrasal frequency effects were found in all eye-movement measures – except for FXP – and among both native and nonnative speakers. Second, controlling for the effects of word-level factors (Word1freq, Word2freq, Word1strokes, Word2strokes) and phrasal frequency, contingency effects were found in FPR and FXC. Specifically, significant effects of contingency emerged when analysing the FXC among both native and nonnative speakers of Chinese, whereas such effects only appeared among nonnative speakers in the first-pass reading of the adverbial sequences. Third, the interaction between phrasal frequency and contingency was found in FPR. However, as evidenced by the significant three-way interaction

Table 9. Mixed-effects logistic model results for fixation probability.

Parameters	Fixed effects			Random effects			
				By subject		By item	
	Estimate	*SE*	*z*	Variance	*SD*	Variance	*SD*
Intercept	1.28	0.37	3.46***	2.24	1.50	0.03	0.17
Phrasalfreq	–0.09	0.14	–0.67	0.03	0.17	–	–
MI (mutual information)	0.09	0.08	1.09	0.01	0.10	–	–
Speaker	3.07	0.59	5.20***	–	–	0.31	0.55
Phrasalfreq × MI	–0.04	0.09	–0.49	0.00	0.05	–	–
Phrasalfreq × Speaker	0.33	0.23	1.43	–	–	–	–
MI × Speaker	–0.17	0.12	–1.39	0.00	0.01	–	–
Phrasalfreq × MI × Speaker	–0.01	0.12	–0.10	–	–	–	–

Note. There are 3,574 observations, where one observation is equal to one FXP measurement for one adverbial sequence read by one participant. Model formula: FXP ~ Phrasalfreq*MI*Speaker + (1 + Phrasalfreq*MI | Subject) + (Speaker | Item). * $p < .05$; ** $p < .01$; *** $p < .001$.

(Phrasalfreq × MI × Speaker), phrasal frequency interacted with contingency only among nonnative speakers. Although higher-frequency sequences were read faster than lower-frequency ones by L2 speakers, such facilitative effect of phrasal frequency was attenuated as the contingency increased. Fourth, both native and nonnative speakers of Chinese were sensitive to statistical properties including phrasal frequency and contingency. The patterns of the sensitivity to phrasal frequency were similar between both groups of language users except in FPR. During the first-pass reading, native speakers processed the adverbial sequences in a surprisingly reversed fashion to that of nonnative speakers; that is, higher-frequency sequences were read at a slower speed than lower-frequency ones. Regarding the sensitivity to contingency, nonnative speakers of Chinese were sensitive to the contingency information of the MWS during both early (FPR) and late processes (FXC), while such sensitivity only appeared in late processes (FXC) for native speakers. Finally, native and nonnative speakers also differed in their degree of sensitivity to phrasal frequency and contingency. Specifically, nonnative speakers of Chinese showed greater degree of sensitivity to the phrasal frequency and contingency of MWS, as revealed by results found in FPR and FXC.

Phrasal frequency effects obtained in MWS in this study confirm the robustness of frequency effects in the following ways. To begin with, it confirms the findings made by previous studies that frequency effects are not limited to the single word level, but extend to MWS as well (Arcara et al., 2012; Arnon and Snider, 2010; Bannard and Matthews, 2008; Conklin and Schmitt, 2008; Durrant and Doherty, 2010; Ellis et al., 2008; Jiang and Nekrasova, 2007; Janssen and Barber, 2012; Siyanova-Chanturia et al., 2011a, 2011b; Sosa and MacFarlane, 2002; Tremblay and Baayen, 2010; Tremblay et al., 2011; Underwood et al., 2004; Wolter and Gyllstad, 2013). Importantly, as a continuous approach was adopted by using stimuli covering a wide frequency range, such findings also suggest that effects of phrasal frequency exist in a continuum (Arnon and Snider,

2010) and are not limited to highly frequent MWS. This also supports the usage-based view of language acquisition and processing in that both highly frequent language formulae and less frequent, more flexible MWS are subject to common statistical learning mechanisms. Lastly, detecting effects of whole-string frequency in nonnative speakers also provides evidence for that sequence-level frequency is also responsible for second language processing (e.g. Wolter and Gyllstad, 2013).

Effects of contingency during the processing of MWS revealed in the current study are also worth mentioning. Most previous studies on the processing of MWS have focused on the role of phrasal frequency, neglecting the role of contingency. Among the research (Ellis et al., 2008; Ellis et al., 2014; Tremblay and Baayen, 2010) that targeted the contingency information of MWS, they are limited by the inadequacy of automaticity of the experimental tasks or techniques. This study adopted the eye-tracking paradigm, which allows more automatic processes than previous studies, hence providing further evidence for the functioning of contingency information during the processing of MWS. From a theoretical perspective, the existence of the effects of phrasal frequency and contingency at the multi-word level among both native and nonnative speakers also supports that both first and second language acquisition and processing are usage-based, and that the processing of MWS are subject to statistical learning mechanisms that include both frequency and contingency.

Combining the independent effects of phrasal frequency/contingency and their interactions as revealed in this study, some light can be shed on the nature of the role of the two statistical properties and their relationship during the processing of MWS. Firstly, both phrasal frequency and contingency of MWS functioned in a facilitative way in terms of reading time. This is supported by the reading patterns of the participants. The more frequent the Chinese sequences were, the faster they were read (in terms of FFD and TRT). Although an impeditive effect of phrasal frequency was found in FPR (estimate = 0.09, SE = 0.04, t = 2.18, p = .029) among native speakers, it is likely that such an effect may not be a true one simply because it could have been brought about by the large sample size (N = 2989) of our data set. Similarly, MWS with higher contingency were also processed faster during the first-pass reading for nonnative speakers. However, in terms of fixation counts, the functioning pattern of phrasal frequency and contingency of MWS seems to differ. For both native and nonnative speakers, higher-frequency sequences required fewer fixations (Phrasalfreq: estimate = −9.77, SE = 4.82, z = −2.03, p = .042), while higher-MI sequences attracted more fixations (MI: estimate = 6.98, SE = 3.34, z = 2.09, p = .037). Given that FXC is a measure that reflects later processes (Paterson et al., 1999; Rayner et al., 1989), this suggests that extra, later stages of processes are needed for more probabilistic adverbial sequences for both native and nonnative speakers. Such processes may include reanalysis of information, integration of information in discourse and recovery from processing difficulties (Paterson et al., 1999; Rayner et al., 1989). Secondly, our results also suggest that phrasal frequency and contingency may function in different time windows during the processing of MWS. As already mentioned, significant effects of phrasal frequency were found in eye movement measures that reflect both early (FFD, FPR) and late (TRT, FXC) processes, indicating that frequency information of the whole word string is activated from the earliest time point during the reading and continue to function until the very end of the processing. By

contrast, contingency is likely to function only in late processes such as reanalysis of information, integration of information in discourse and recovery from processing difficulties during the processing of MWS. Although effects of contingency were detected by an early measure (i.e. FPR) for nonnative speakers, we suppose such effects were likely to result from early processes such as familiarity check (Reichle et al., 1998; Roberts and Siyanova-Chanturia, 2013) due to the property of the stimuli used in this study. In this research, familiarity ratings of the MWS received from the nonnative speakers were not satisfactorily high (Table 3), thus familiarity check could be necessary. Since earlier processes such as familiarity check driven by contingency happened in the same time window as those driven by phrasal frequency, this also explains why the effects of phrasal frequency were attenuated during the first-pass reading as MI values increased for nonnative speakers of Chinese.

In terms of the statistical sensitivity to the phrasal frequency and contingency of MWS among native and nonnative speakers, several interesting findings were revealed. First, the results suggest that both native and nonnative speakers are sensitive to the phrasal frequency of MWS. Native speakers' sensitivity to the whole-string frequency information has been supported by mounting evidence, yet whether nonnative speakers (including advanced nonnative speakers) are sensitive to the phrasal frequency of MWS is still under debate. Our research provides firm evidence in favor of the existence of the sensitivity to phrasal frequency among advanced second language learners, in that significant or marginally significant effects of phrasal frequency were found in almost every eye movement measures (except FXP). Second, the results also suggest that advanced nonnative speakers can be sensitive to the contingency information of MWS just like native speakers. Current studies (e.g. Ellis et al., 2008; Tremblay and Baayen, 2010) have found significant effects of contingency only among native speakers and claimed that L2 speakers (including advanced L2 speakers) are still not tuned to the contingency information due to their limited sampling of the target language (Ellis et al., 2008). However, our findings indicate that advanced nonnative speakers do have statistical knowledge about the contingency information of MWS, and they do take advantage of such knowledge to boost their processing of L2 MWS. Importantly, such findings were made by using the eye-tracking technique that allows more automatic processes than many other experimental tasks used by previous studies (e.g. Ellis et al., 2008). Moreover, given that the average length of the formal L2 instruction of our participants was less than four and a half years, such results also suggest that the tuning of second language statistics may start from earlier stages than what Ellis et al. (2008) suggested. First language learners are found to be sensitive to the underlying statistical properties since infancy (Saffran et al., 1996); hence, more studies are needed to explore the timing of statistical tuning in second language acquisition.

Furthermore, this study also indicates that native and nonnative speakers may differ in their degrees of statistical sensitivity, and advanced nonnative speakers may be more sensitive to statistical regularities including the phrasal frequency and contingency of MWS than native speakers. This is supported by nonnative speakers' significant greater magnitude of the effects of phrasal frequency (in FPR and FXC) and contingency (in

FPR and FXC) as evidenced by the significant interaction effects between Phrasalfreq/MI and Speaker (see results reported in the previous section). More studies are needed to test such sensitivity patterns, yet the difference in the degrees of sensitivity to the phrasal frequency and contingency of MWS between native and nonnative speakers is possible especially given the findings made by current research. For example, Duyck et al. (2008) directly compared effects of word frequency in the first and second language using two lexical decision tasks and they found that Dutch-English bilinguals showed a considerably larger frequency effect in their second language. The possibility that nonnative speakers may be more sensitive to the underlying statistical information is also supported by the power law of practice in language acquisition (DeKeyser, 2007). Learning effects follow a logarithmic fashion, and progressively smaller learning effects take places with each occurrence of language use or practice. Therefore, given that nonnative speakers are still at earlier stages of using or practicing their second language, stronger effects of phrasal frequency and contingency should be expected, leading to larger magnitude of reduction in reaction time as phrasal frequency or contingency increases than that in native speakers.

Finally, revealing that nonnative speakers were sensitive to both phrasal frequency and contingency of MWS as native speakers were, it suggests that native and nonnative speakers are likely to share the same general statistical learning mechanism when processing MWS. As reviewed in the beginning of this article, statistical sensitivities are retained from childhood (e.g. Gomez and Gerken, 2000; Saffran et al., 1996) to adulthood (e.g. Frank et al., 2010; Zuhurudeen and Huang, 2016), and from first language acquisition (e.g. Saffran et al., 1996) to second language acquisition (Frost et al., 2013; Hamrick, 2014). However, it remains unclear whether first and second language users are sensitive to the same statistical properties. This is especially true when one considers the heated debate about the critical period of language acquisition and the native-likeness of the ultimate L2 attainment (e.g. DeKeyser and Larson-Hall, 2005; Granena and Long, 2013; Johnson &Newport, 1989; Long, 2005), or the dispute over whether implicit or explicit learning dominates second language acquisition, and whether native and nonnative speakers fundamentally differ in terms of their way of processing the language (e.g. Hulstijn, 2005). Such debates raise the possibility that second language learners may be maturationally constrained and that they may differ from native speakers in their way of processing the second language. Despite such concerns, our findings support not only the idea the statistical sensitivity is retained in nonnative speakers, but also that second language learners show sensitivities to different kinds of statistical information in a similar fashion as native speakers, at least in the case of the phrasal frequency and contingency of MWS.

Acknowledgements

We would like to express our gratitude to the anonymous *Second Language Research* reviewers, and to the editor Alice Foucart, for their insightful criticism and generous feedback. We would like to thank Xingshan Li for allowing us to use the eye-tracking lab and the equipment. We are also grateful to Steven Ross, Nicoo Cooper, Philip Blue and Andrew Cavanaugh for their proofreading of this article. We thank Yucheng Liu for his assistance in data collection.

Declaration of conflicting interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the China National Social Science Fund (grant number 12BYY059).

Notes

1. All sample MWS in this article were extracted from the British National Corpus.
2. LogitABCD is a contingency measure used by Tremblay and Baayen (2010). It refers to the log probability of obtaining the word D given the sequence of ABC. The formula used by the authors to compute LogitABCD is: LogitABCD = log(FrequencyABCD / (FrequencyABC – FrequencyABCD + 1)).
3. The Chinese sentence was spaced to improve the readability, followed by the Chinese pinyin annotation, then by its word-by-word English translation and the English translation.
4. When computing MI scores, frequency counts of homonyms and homographs sharing the same Chinese character with the adverbial use of the constituent words were removed.
5. Cutoff points for phrasal frequency and MI are arbitrary. The only reason to use them is to divide candidate materials into different subgroups.
6. Formula: Dependent variable ~ Phrasalfreq*MI*Speaker + (1 + Phrasalfreq*MI| Subject) + (Speaker| Item)
7. The formula is 2 * (1 – pt(abs(X), Y – Z)). X is the *t* value, Y is the number of observations, and Z is the number of fixed effect parameters.

References

Academia Sinica (2015) *Sinica balanced Modern Chinese corpus*. Taipei: Academia Sinica. Available at: http://asbc.iis.sinica.edu.tw/index_range.htm (accessed April 2017).

Arcara G, Lacaita G, Mattaloni E et al. (2012) Is 'hit and run' a single word? The processing of irreversible binomials in neglect dyslexia. *Frontiers in Psychology* 3: 1–11.

Arnon I and Snider N (2010) More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62: 67–82.

Baayen RH (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Balota D, Cortese M, Sergent-Marshall S, Spieler D, and Yap M (2004) Visual word recognition for single-syllable words. *Journal of Experimental Psychology: General* 133: 283–316.

Bannard C and Matthews D (2008) Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science* 19: 241–48.

Barr DJ, Levy R, Scheepers C, and Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68: 255–78.

Bates D, Maechler M, Bolker B, and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.

Biber D, Johansson S, Leech G, Conrad S, and Finegan E (1999) *Longman grammar of spoken and written English*. Harlow: Longman.

Bod R (1998) *Beyond grammar: An experience-based theory of language*. Stanford, CA: CSLI.

Bod R (2006) Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review* 23: 291–320.

Bybee J (1998) The emergent lexicon. *Chicago Linguistic Society* 34: 421–35.
Center for Chinese Linguistics, Peking University (2015) *CCL Modern Chinese corpus*. Beijin: Center for Chinese Linguistics. Available at: http://ccl.pku.edu.cn:8080/ccl_corpus (accessed April 2017).
Christiansen MH and Chater N (1999) Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23: 157–205.
Church K, Gale W, Hanks P, and Hindle D (1991) Using statistics in lexical analysis. In: Zernik U (ed.) *Lexical acquisition: Exploiting on-line resources to build a lexicon*. Hillsdale, NJ: Lawrence Erlbaum, pp. 115–64.
Conklin K and Schmitt N (2008) Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29: 72–89.
Conway CM, Bauernschmidt A, Huang S, and Pisoni DB (2010) Implicit statistical learning in language processing: Word predictability is the key. *Cognition* 114: 356–71.
DeKeyser R (2007) Skill acquisition theory. In: VanPatten B and Williams J (eds) *Theories in second language acquisition: An introduction*. 1st edition. New York: Routledge, pp. 97–113.
DeKeyser R and Larson-Hall J (2005) What does the critical period really mean? In: Kroll JF and de Groot AMB (eds) *Handbook of bilingualism: Psycholinguistic approaches*. Oxford: Oxford University Press, pp. 89–108.
Diessel H (2007) Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25: 104–23.
Durrant P and Doherty A (2010) Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic theory* 6: 125–55.
Duyck W, Vanderelst D, Desmet T, and Hartsuiker RJ (2008) The frequency effect in second-language visual word recognition. *Psychonomic Bulletin and Review* 15: 850–55.
Ellis NC (2002) Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24: 143–88.
Ellis NC (2003) Constructions, chunking, and connectionism: The emergence of second language structure. In: Doughty CJ and Long MH (eds) *The handbook of second language acquisition*. Oxford: Blackwell, pp. 63–103.
Ellis NC (2006a) Language acquisition as rational contingency learning. *Applied Linguistics* 27: 1–24.
Ellis NC (2006b) Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics* 27: 164–94.
Ellis NC (2008) Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In: Robinson P and Ellis NC (eds) *Handbook of cognitive linguistics and second language acquisition*. New York and London: Routledge, pp. 372–405.
Ellis NC, O'Donnell MB, and Romer U (2014) The processing of verb–argument constructions is sensitive to form, function, frequency, contingency, and prototypicality. *Cognitive Linguistics* 25: 55–98.
Ellis NC, Simpson-Vlach R, and Maynard C (2008) Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly* 42: 375–96.
Elman JL (1990) Finding structure in time. *Cognitive Science* 14: 179–211.
Engbert R, Nuthmann A, Richter EM, and Kliegl R (2005) SWIFT: A dynamical model of saccade generation during reading. *Psychological Review* 112: 777–813.
Erickson LC and Thiessen ED (2015) Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review* 37: 66–108.

Erman B and Warren B (2000) The idiom principle and the open choice principle. *Text* 20: 29–62.
Fang Q (2012) Xiandai hanyu fuci lianyong pinlv kaocha [The frequency distribution of modern Chinese adverbial sequences]. *Hanyu Xuebao* 39: 87–94.
Fine AB and Jaeger FT (2013) Evidence for implicit learning in syntactic comprehension. *Cognitive Science* 37: 578–91.
Fox J (2003) Effect displays in R for generalized linear models. *Journal of Statistical Software* 8: 1–27.
Frank MC, Goldwater S, Griffiths TL, and Tenenbaum JB (2010) Modeling human performance in statistical word segmentation. *Cognition* 117: 107–25.
Frost R, Siegelman N, Narkiss A, and Afek L. What predicts successful literacy acquisition in a second language? *Psychological Science* 24: 1243–52.
Frost R, Amstrong BC, Siegelman N, and Christainsen MH (2015) Domain generality vs. modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences* 19: 117–25.
Goldberg AE (1995) *Constructions: A construction grammar approach to argument structure*. Chicago, IL: Chicago University Press.
Goldberg AE (2006) *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Gomez RL and Gerken L (2000) Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences* 4: 178–86.
Granena G and Long MH (2013) Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research* 29: 311–43.
Gregory M, Raymond W, Bell A, Fosler-Lussier E, and Jurafsky D (1999) The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society* 35: 151–66.
Gries ST (2010) Useful statistics for corpus linguistics. In: Sánchez A and Almela M (eds) *A mosaic of corpus linguistics: Selected approaches*. Frankfurt: Peter Lang, pp. 269–91.
Gries ST and Ellis NC (2015) Statistical measures for usage-based linguistics. *Language Learning* 65: 228–55.
Hamrick P (2014) A role for chunk formation in statistical learning of second language syntax. *Language Learning* 64: 247–78.
Hulstijn J (2005) Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction. *Studies in Second Language Acquisition* 27: 129–40.
Janssen N and Barber HA (2012) Phrasal frequency effects in language production. *PloS One* 7: 1–11.
Jescheniak JD and Levelt W (1994) Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20: 824–43.
Jiang N and Nekrasova TM (2007) The processing of formulaic sequences by second language speakers. *Modern Language Journal* 91: 433–45.
Johnson JS and Newport EL (1989) Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21: 60–99.
Jurafsky D (2003) Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In: Bod R, Hay J, and Jannedy S (eds) *Probabilistic linguistics*. Cambridge, MA: MIT Press, pp. 39–96.
Jurafsky D, Bell A, Gregory M, and Raymond W (2001) Probabilistic relations between words: Evidence from reduction in lexical production. In: Bybee J and Hopper P (eds) *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp. 229–54.
Langacker RW (1987) *Foundations of cognitive grammar*. Stanford, CA: Stanford University Press.

Li C (2010) Shou lianyong fuci yueshu de VP yukuai de tiqu yu yanjiu [The extraction of VP constructions modified by adverbial sequences]. Unpublished Master's thesis, Peking University, Beijing, China.

Li X, Liu P, Wei W, Bicknell K, and Rayner K (2014) Reading is fundamentally similar across disparate writing systems: A systematic characterization of how words and characters influence eye movements in Chinese reading. *Journal of Experimental Psychology: General* 143: 895–913.

Liversedge SP, Drieghe D, Zang C, Zhang M, Bai X, and Yan G (2014) The effect of visual complexity and word frequency on eye movements during Chinese reading. *Visual Cognition* 22: 441–57.

Long M (2005) Problems with supposed counter-evidence to the critical period hypothesis. *International Review of Applied Linguistics in Language Teaching (IRAL)* 43: 287–317.

Ma G and Li X (2015) How character complexity modulates eye movement control in Chinese reading. *Reading and Writing: An Interdisciplinary Journal* 28: 747–61.

MacWhinney B (1998) Models of the emergence of language. *Annual Review of Psychology* 49: 199–227.

Maye J, Weiss DJ, and Aslin RN (2008) Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science* 11: 122–34.

McDonald SA and Shillcock RC (2003) Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research* 43: 1735–51.

National Committee of Language and Script (2015a) *Classified word frequency list of Modern Chinese*. Beijing: National Committee of Language and Script. Available at: http://www.cncorpus.org/resources/CorpusWordPOSlist.xls (accessed April 2017).

National Committee of Language and Script (2015a) *Modern Chinese corpus*. Beijing: National Committee of Language and Script. Available at: http://www.cncorpus.org/Resources.aspx (accessed April 2017).

Paterson K, Liversedge S, and Underwood G (1999) The influence of focus operators on syntactic processing of short relative clause sentences. *The Quarterly Journal of Experimental Psychology* 52: 717–37.

Pawley A and Syder FH (1983) Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In: Richards JC and Schmidt RW (eds) *Language and communication*. London: Longman, pp. 191–226.

Pierrehumbert JB (2001) Exemplar dynamics: Word frequency, lenition and contrast. In: Bybee J and Hopper P (eds) *Frequency effects and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp. 137–58.

Pinker S (1999) *Words and rules: The ingredients of language*. New York: Basic Books.

Pinker S and Ullman MT (2002) The past and future of the past tense. *Trends in Cognitive Sciences* 6: 456–63.

Portin M, Lehtonen M, Harrer G, Wande E, Niemi J, and Laine M (2008) L1 effects on the processing of inflected nouns in L2. *Acta Psychologica* 128: 452–45.

R Core Team (2015). R: A language and environment for statistical computing. *Vienna: R Foundation for Statistical Computing*. Available at: http://www.R-project.org (accessed April 2017).

Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124: 372–422.

Rayner K, Li X, Juhasz BJ, and Yan G (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin and Review* 12: 1089–93.

Rayner K, Sereno SC, Morris RK, Schmauder AR, and Clifton C (1989) Eye movements and online comprehension processes. *Language and Cognitive Processes* 4: 21–49.

Rebuschat P (2013) Statistical learning. In: Robinson P (ed.) *The Routledge encyclopedia of second language acquisition*. London: Routledge, pp. 612–15.

Reichle ED, Pollatsek A, Fisher DL, and Rayner K (1998) Toward a model of eye movement control in reading. *Psychological Review* 105: 125–57.

Roberts L and Siyanova-Chanturia A (2013) Using eye-tracking to investigate topics in L2 acquisition and L2 sentence and discourse processing. *Studies in Second Language Acquisition* 35: 213–35.

Rumelhart D and McClelland J (1986) On learning the past tenses of English verbs. In: Rumelhart D and McClelland J (eds) *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press, pp. 216–71.

Saffran JR (2003) Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science* 12: 110–14.

Saffran JR, Aslin R N, and Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274: 1926–28.

Schmauder AR, Morris RK, and Poynor DV (2002) Lexical processing and text integration of function and content words: Evidence from priming and eye fixations. *Memory and Cognition* 28: 1098–1108.

Schmidt JR (2012) Human contingency learning. In: Seal NM (ed.) *Encyclopedia of the Sciences of Learning*. New York: Springer, pp. 1455–56.

Segui J, Mehler M, Frauenfelder U, and Morton J (1982) The word frequency effect and lexical access. *Neuropsychologia* 20: 615–27.

Simpson-Vlach R and Ellis NC (2010) An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31: 487–512.

Siyanova-Chanturia A, Conklin K, and Schmitt N (2011a) Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research* 27: 251–72.

Siyanova-Chanturia A, Conklin K, and van Heuven W (2011b) Seeing a phrase 'time and again' matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37: 776–84.

Shanks DR (1995) *The psychology of associative learning*. New York: Cambridge University Press.

Sosa A and MacFarlane J (2002) Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language* 83: 227–36.

Swingley D (2005) Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology* 50: 86–132.

Thiessen ED and Saffran JR (2003) When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month–old infants. *Developmental Psychology* 39: 706–16.

Thompson SP and Newport EL (2007) Statistical learning of syntax: The role of transitional probability. *Language Learning and Development* 3: 1–42.

Tomasello M (2001) First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics* 11: 61–82.

Tomasello M (2003) *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tremblay A and Baayen RH (2010) Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In: Wood D (ed.) *Perspectives on formulaic language: Acquisition and communication*. London: Continuum, pp. 151–73.

Tremblay A, Derwing B, Libben G, and Westbury C (2011) Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning* 61: 569–613.

Underwood G, Schmitt N, and Galpin A (2004) The eyes have it: An eye-movement study into the processing of formulaic sequences. In: Schmitt N (ed.) *Formulaic sequences: acquisition, processing and use*. Amsterdam: John Benjamins, pp. 153–72.

Wolter B and Gyllstad H (2011) Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics* 32: 430–49.

Wolter B and Gyllstad H (2013) Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition* 35: 451–82.

Wood D (2002) Formulaic language in acquisition and production: Implications for teaching. *TESL Canada Journal* 20: 111–18.

Wray A (2002) *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zang C, Liversedge S, Bai X, and Yan G (2011) Eye movements during Chinese reading. In: Liversedge S, Gilchrist I, and Everling S (eds) *The Oxford handbook of eye movements*. Oxford: Oxford University Press, pp. 961–78.

Zuhurudeen F and Huang Y (2016) Effects of statistical learning on the acquisition of grammatical categories through Qur'anic memorization: A natural experiment. *Cognition* 148: 79–84.

Appendix 1. The monosyllabic Chinese adverbs that constructed the experimental multiword sequences (MWS).

Adverb	Chinese pinyin	Normalized frequency	Literal translation
别	bié	194	do not
并	bìng	1,022	also
不	bù	5,023	not/no
才	cái	565	just
曾	céng	380	once
倒	dào	223	actually
都	dōu	1,522	both/all
更	gèng	750	more
还	hái	1,608	still/yet
很	hěn	974	very
将	jiāng	86	will
就	jiù	2,356	then/only
没	méi	1,309	have/has not
却	què	511	but
仍	réng	289	still
太	tài	470	too
挺	tǐng	46	rather
也	yě	2,302	also
已	yǐ	1,370	already
又	yǒu	1,094	again
再	zài	671	once more
真	zhēn	430	really
只	zhǐ	1,048	only
最	zuì	855	the most/-est

Notes. Raw phrasal frequencies were obtained from the CCL Corpus (size: 300 million words). Normalized phrasal frequency was computed using the following formula:
$Normalized\ phrase\ frequency = \left(Raw\ phrase\ frequency \times 1,000,000\right) / 300,000,000$.