Research Article

Xiaolong Lu*

Enhancing the processing advantage: two psycholinguistic investigations of formulaic expressions in Chinese as a second language

https://doi.org/10.1515/iral-2023-0262
Received October 16, 2023; accepted June 3, 2024; published online July 22, 2024

Abstract: The Holistic Hypothesis asserts that formulaic expressions (FEs) are processed more rapidly than non-formulaic expressions (non-FE items) by both native speakers (NSs) and L2 learners of English. This study utilized an acceptability judgment task and a self-paced reading task to investigate the online processing of FEs and non-FE items among bilingual speakers (L1 English, L2 Chinese) in both contextual and non-contextual conditions. Meanwhile, a familiarity rating task was employed to measure whether there is a familiarity effect in item processing. The results consistently provided support for the Holistic Hypothesis, indicating that learners of Chinese at each level exhibited faster processing of FEs compared to non-FE items, regardless of the presence or absence of context. However, the influence of item familiarity, rather than the proficiency effect, contributed to the improvement of L2 learners’ eventual processing abilities. Distinct patterns also emerged when comparing data from NSs and L2 learners of Chinese, highlighting L2 learners’ more pronounced processing advantage, characterized by faster response times (RTs) to FEs compared to non-FE items. Through an analysis of Chinese L2 data, this study sheds light on the interplay between the usage based approach and chunking within the cognitive approach to L2 learning.

Keywords: formulaic expressions; holistic hypothesis; proficiency; familiarity; L2 Chinese

1 Introduction

Formulaic expressions (FEs), also referred to as chunks or multi-word sequences, have been extensively examined in the field of language processing. It has been

*Corresponding author: Xiaolong Lu, East Asian Studies, The University of Arizona, 825E 5th Street, 85721-0001 Tucson, AZ, USA, E-mail: charmander@arizona.edu

Open Access. © 2024 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
recognized that our brains store representations of individual words in addition to FEs, which are frequently occurring sequences that can be readily retrieved from long-term memory (Conklin and Schmitt 2012). FEs serve as linguistic units that bridge the gap between lexicon and grammar, and they are commonly employed in both spoken and written languages (Wray 1999). Studies in psycholinguistics (e.g., McCauley and Christiansen 2014) have provided support for the idea that learning and using FEs form the foundation of human comprehension and production processes. This pertains to the identification of multiword units (through distributional learning) from natural language input by children learning their first language (L1) (e.g., Arnon and Clark 2011; Bannard and Matthews 2008), as well as by adults learning a second language (L2) (e.g., Arnon and Snider 2010; Jolsavi et al. 2013).

The processing of FEs as integral building blocks of language (Arnon and Christiansen 2017) reflects how words are grouped together to form larger-than-word units in the mental lexicon of language speakers. Previous studies employed multiple experimental paradigms, such as a phrasal-decision task (e.g., Arnon and Snider 2010; Hernández et al. 2016; Supasiraprapa 2019), self-paced reading (e.g., Kim and Kim 2012; Shantz 2017; Tremblay et al. 2011), eye-tracking technique (e.g., Piumégé et al. 2021; Pulido 2021; Yi et al. 2017), and event-related brain potentials (e.g., Paulmann et al. 2015; Siyanova-Chanturia 2013; Tremblay and Baayen 2010). What they found is that both NSs and L2 learners are sensitive to these distributional properties when processing FEs from the input, suggesting that the processing of FEs as multiword units is subject to common mechanisms (e.g., chunking) of statistical learning in both L1 and L2 acquisition.

One of the most important factors regarding FE processing is its holistic nature. The Holistic Hypothesis (Schmitt et al. 2004) argues that FEs (or recurrent word clusters) are holistically stored, easily discerned and retrieved in human minds, thus “allowing language users to be more fluent while freeing up cognitive resources for other language processes (p. 128)”. The hypothesis is argued to represent the nature of FEs in language processing, showing the processing of FEs and non-FE items might be different for both NSs and L2 learners (Conklin and Schmitt 2012). Subsequent studies (e.g., Conklin and Schmitt 2012; Jiang and Nekrasova 2007) largely extended the investigation of the holistic account to the L2 population and found that both NSs and advanced-level L2 learners responded to FEs more rapidly than to regular, non-formulaic phrases (non-FE items), which required syntactic analysis.

The study aims to investigate and validate the holistic nature of FEs in visual processing, with a specific emphasis on the understudied individuals who have Chinese as L2 and English as L1. The subsequent section will comprehensively review various factors involved in the processing of FEs to situate the present research.
2 Literature review

There are multiple factors contributing to the visual processing of FEs in the L2 context. Previous studies generally examined the role of frequency (e.g., Ellis et al. 2008; Schmitt et al. 2004; Wolter and Gyllstad 2013; Wolter and Yamashita 2018; Wray 2012), familiarity (e.g., Carrol et al. 2016; Carrol and Conklin 2017; Tabossi et al. 2009; Van Lancker Sidtis 2012), proficiency (e.g., Fang and Zhang 2021; Nekrasova 2009; Sonbul 2015; Vilkaïtė and Schmitt 2019; Wolter and Yamashita 2018), congruency (word-for-word translations between L1 and L2, e.g., Chen 2024; Ding and Reynolds 2019; Gyllstad and Wolter 2016; Sonbul and El-Dakhs 2020; Wolter and Gyllstad 2011, 2013; Yamashita and Jiang 2010), and holistic nature (e.g., Conklin and Schmitt 2008; Jiang and Nekrasova 2007; Kim and Kim 2012; Tremblay et al. 2011) on the processing of FEs. The following section will provide a more detailed review of relevant factors for this study.

2.1 Familiarity effect

Carrol et al. (2016) conducted an eye-tracking while reading study to examine how advanced Swedish (L1) learners of English (L2) processed English idioms. To do so, they asked their participants to read English-only idioms (e.g., bite the bullet), Swedish-only idioms (e.g., born in the hall), and congruent idioms (e.g., break the ice), along with the corresponding control phrases (e.g., grab the bullet) as they were embedded in context-neutral sentences, and then to rate their familiarity with each test item. They found that familiarity showed consistent effects on both NSs and learners of English. For native English speakers, better-known English-only and congruent idioms were easier to understand, demonstrated by a higher likelihood of skipping the final word, and shorter total reading time. For L2 learners, L1 familiarity seemed to facilitate their processing of congruent and Swedish-only idioms, whereas L2 familiarity played a bigger role in the processing of English-only idioms where no L1 knowledge existed. In another study, Carrol and Conklin (2017) designed eye-tracking experiments to investigate the reading of word-for-word translated Chinese idioms and English idioms by comparing Chinese-speaking learners of English and monolingual English NSs. In their first experiment, they designed English idioms (e.g., a chip off the old block) and controls (e.g., a chip off the old wall), as well as English-translated Chinese idioms (e.g., half-believe-half-doubt, meaning ‘dubious’) and controls (e.g., half-believe-half-judge), with the goal of examining whether the preceding words were sufficient to facilitate lexical access to the final word for each item. They gauged familiarity by requesting English and Chinese native speakers to
rate all the items presented in English. Chinese native speakers not only rated all these English items but also Chinese idioms presented in their original Chinese characters. The result showed that there was no significant familiarity effect on the reading of Chinese or English idioms by the L2 learners, while a familiarity effect facilitated the likelihood of skipping by English NSs.

Overall, the findings of the two eye-tracking studies indicated inconsistent evidence regarding the extent to which familiarity facilitates L2 learners’ processing. Consequently, the effect of familiarity on the processing of FEs remains inconclusive, unless multiple experimental paradigms are utilized to triangulate its impact.

### 2.2 Proficiency effect

Nekrasova (2009) conducted a gap-filling task and a dictation task to test whether NS and L2 English speakers displayed any knowledge of lexical bundles as holistic units (FEs), and whether their knowledge was affected by the discourse function of FEs. The first task was employed to measure speakers’ precise knowledge of the constituent words in FEs, and the second task was used to gauge the participants’ knowledge of the holistic forms of the FEs. She compared intermediate and advanced learners of English and English NSs’ knowledge of two FE types: discourse-organizing bundles (e.g., if you look at) and referential bundles (e.g., in the middle of). Her finding from the gap-filling task showed that intermediate learners scored significantly lower than the other two groups, whereas there was no significant difference between advanced learners and NSs. This indicated that the proficiency effect played a role in learners’ accurate production and comprehension of FEs. On the dictation task, similarly, advanced learners outperformed intermediate learners in the amount and accuracy of the FEs that they recalled. Moreover, the comparisons of the two FE types showed that all three groups acquired discourse organizers better than referential bundles on the two tasks, indicating that the speakers’ knowledge of FEs might be affected by linguistic register and discourse function. The author concluded that the proficiency effect in the processing of FEs was due to the fact that higher proficiency L2 learners might have acquired greater lexical knowledge, which could lead to their enhanced knowledge of FEs. As previously mentioned, Wolter and Yamashita (2018) also tested the effect of proficiency on L2 collocational processing. They employed a phrase acceptability judgment task to compare how intermediate, advanced, and native speakers of English processed congruent and incongruent collocations. In contrast to NSs, they found that L2 learners processed congruent collocations significantly faster than English-only collocations. The proficiency effect was tested in the sense that learners’ increased proficiency enabled them to undergo a shift from sensitivity to word frequency to sensitivity to collocational frequency,
thus revealing a development toward nativeness. Based on this developmental pattern, the authors challenged previous claims that L2 learners may process FEs differently than NSs, arguing that the development of learners’ ability to process FEs was associated with language proficiency.

In conclusion, the two studies employed distinct experiments to highlight the influence of L2 proficiency on FE processing, revealing that L2 learners with higher proficiency levels demonstrate a greater ability to retain FEs compared to those with lower levels.

2.3 Holistic nature of FEs

The holistic nature of FEs is closely related to a low degree of compositionality, particularly in idiomatic expressions. According to the principle of compositionality (Pelletier 1994), the meaning of a linguistic unit depends on the meanings of its parts and their syntactic combination. It is understood that certain FEs, such as idioms, exhibit low compositionality, signifying their conventional meanings and frozen structures (characteristic of holistic storage). In contrast, other FEs, like collocations, display high compositionality, revealing their literal meanings and analyzable structures. It is therefore important to investigate whether highly compositional FEs are processed differently from those with low compositionality. Conklin and Schmitt (2008) used a self-paced reading paradigm to investigate whether an FE which required a figurative interpretation with a low compositionality (e.g., a breath of fresh air ‘a new approach’) was processed faster than the same one with a high compositionality (e.g., a breath of fresh air ‘breathing clean air outside’). Their results showed that NSs and L2 learners processed both types equally fast, suggesting that compositionality might not be an important factor. However, the opposite finding was obtained in an eye-tracking experiment carried out by Siyanova-Chanturia et al. (2011). They measured how participants read passages containing idioms, thus low in compositionality (e.g., left a bad taste in my mouth) and corresponding high-compositionality control phrases (e.g., the bad taste left in his mouth). Their various eye-tracking measurements indicated that NSs processed the idioms significantly faster than the controls, while the reverse pattern was found for L2 learners. Their results were interpreted as showing that only the NSs hold an advantage in processing FEs that were low in compositionality (i.e., idioms), while L2 learners did not.

In summary, the two studies showed that the empirical evidence on the role noncompositionality or holistic storage might play in the processing of FEs by NSs and L2 learners is still contrasting. Part of this contrast could be attributed to the attention directed toward idioms as distinctive forms of FEs, which were examined using various experimental paradigms.
2.4 Identifying and processing of FEs in L2 Chinese

The classic definition of FEs comes from Wray (2002: 9), showing that an FE is “a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.” Nevertheless, Wang (2020) raised doubts regarding its suitability for the Chinese language, citing concerns about Wray’s incorporation of various elements such as morphemes. Additionally, she observed distinctions between Chinese and English, highlighting Chinese as a character-based language lacking inter-word spacing. Drawing from past definitions of Chinese FEs, she concluded that a multi-word expression (or FE) constituted a prefabricated language communication unit, comprising two or more words or morphemes. She argued that the prefabricated nature of FEs mirrored the integrity of information storage and retrieval in human language processing. This aligns with the holistic characteristics of FEs in this study. Wang (2020: 46–48) also provided an elaborate categorization for Chinese FEs, including idioms (set phrases, institutionalized expressions, and other types of idioms), polite formulas, parentheses, conventionalized expressions, high-frequency collocations, frame structures (phrase frames, four-character frames), classifier phrases (nominal and verbal classifier phrases). This study selected and examined particular types of FEs to accommodate the learning conditions of L2 participants, considering their limited vocabulary sizes and the use of novice-level textbooks. These FEs include high-frequency collocations (e.g., 太好了 great, 有事儿 have something to do) and classifier phrases (e.g., 一张纸 a paper, 一封信 a letter).

With regard to the empirical studies on FEs in Chinese as a second language, there is still a paucity of investigation into how FEs in Chinese are visually processed in experimental settings. Zheng et al. (2016) employed an acceptability judgment task and a self-paced reading task to examine the processing of three- and four-character FEs and non-FE items, with and without contexts, by NSs and L2 learners of Chinese. Their result showed the processing advantage of these Chinese sequences was context-sensitive. While NSs and L2 learners demonstrated significant processing advantages of FEs without contexts, the processing advantage was reduced when contexts were provided. By using the eye-tracking while reading paradigm, Yi et al. (2017) explored the role of phrasal frequency and contingency (the occurrence probability of each component in FEs) on the on-line processing of disyllabic Chinese adverbial sequences. Their findings suggested that both NSs and learners of Chinese were sensitive to the phrasal frequency and contingency of adverbial sequences in contexts. Another study carried out by Zheng et al. (2022) triangulated the on-line processing data, using an acceptability judgment task with and without think-aloud...
protocols to investigate Chinese NSs and L2 learners’ processing of idioms (a type of FEs) and matched non-idioms (non-FE items). Their experiment with and without think-aloud showed similar patterns. For NSs, FEs were processed faster than non-FE items regardless of length, while L2 learners processed three-character items faster than four-character ones regardless of item type. They concluded that FEs and non-FE items were processed differently for NSs and L2 learners. These studies indicated that the processing advantage of FEs in L2 Chinese depends on the frequency and the type of item being tested.

2.5 Research gaps and questions

The reviewed studies support the Holistic Hypothesis in the L2 population across proficiency levels and investigate the impact of frequency and compositionality on processing FEs. However, the role of frequency, especially derived from L1 corpus data, may not serve as a direct indicator of L2 learners’ fast processing and improved accuracy. This is because the context in which L1 and L2 learners receive input can vary, as noted by Bardel and Falk (2021). Moreover, it has long been asserted that experiential familiarity, gauged through subjective ratings, can serve as a more accurate measure than printed word frequency. This method takes into consideration various encounters with a particular lexical item, encompassing exposure through language production, as well as familiarity with auditory, visual, and written forms (e.g., Connine et al. 1990; Gernsbacher 1984). Familiarity, in this sense, is closely intertwined with frequency, but its direct association with L2 learners’ test performances becomes more valuable when considering it as a factor in online processing. The role of familiarity in explaining the processing differences between FEs and non-FE items in the Chinese L2 context has not been thoroughly investigated. To address this, familiarity ratings are suggested to be included as a subjective measurement (e.g., Carrol et al. 2016). Additionally, previous studies using different experimental methods have yielded inconsistent results, suggesting a need for methodological triangulation. Following Zheng et al. (2016), conducting both a self-paced reading task and an acceptability judgment task will provide converging evidence on holistic processing by examining how learners of Chinese process FEs with and without context.

Previous studies (e.g., Conklin and Schmitt 2012; Jiang and Nekrasova 2007) have investigated the Holistic Hypothesis in non-native speakers and found that NSs and advanced-level L2 learners process FEs faster than non-FE items. However, they overlooked the possibility that lower-level learners, with less exposure to the target language, might not exhibit the same processing advantage. It is hypothesized that lower-level learners might process FEs similarly to non-FE items, resulting in similar
processing time and accuracy. Therefore, it remains unclear whether the processing advantage of FEs extends to lower proficiency levels. Collecting data from these learners will complement existing findings that primarily focus on higher-level L2 learners. As we see, there have only been a handful of psycholinguistic studies (Yi et al. 2017; Zheng et al. 2022) examining FE processing in Chinese as a non-native language. To fill this void in the literature, the study aims to reexamine L2 learning of FEs from a psycholinguistic perspective and further investigate the nature of Chinese FEs by centering on learners of Chinese across different proficiency levels. The present study addresses three research questions (RQs). The first question aims to test whether L2 learners generally have a processing advantage for Chinese FEs (over non-FE items) in two different experiments. The second question explores the effects of familiarity and proficiency on learners’ visual processing of Chinese FEs. The third question investigates the data from the two experiments by comparing the FE processing patterns of learners with those of native Chinese speakers.

**RQ1:** Do learners of Chinese at different proficiency levels show a processing advantage for FEs in the acceptability judgment task (without context) and the self-paced reading task (with context)?

**RQ2:** Do factors such as familiarity and L2 proficiency play a role in L2 learners’ online visual processing of FEs in Chinese?

**RQ3:** Do learners and native speakers of Chinese process FEs in a similar or different way?

### 3 Methods

#### 3.1 Participants

Ninety-four participants contributed to this study, including 34 learners of Chinese (novice to intermediate\(^1\)), 29 learners of Chinese (advanced), and 31 native speakers.

---

\(^1\) The evaluation of L2 learners’ proficiency commenced by considering their class levels and teacher assessments. Due to the unique teaching conditions at the university, which encompassed variables like class enrollments and teaching progress, students at the novice level (CHN 101, 102) were combined with a limited number of intermediate-level students (CHN 201, 202). Consequently, these merged learner groups were designated as beginning-intermediate or lower-level participants under the guidance of teachers. To enhance data collection, advanced-level Chinese learners were chosen not only from CHN 401 classrooms but also from the critical language program. The ILR scale level 2
of Chinese as the control group. All of the learners were recruited from a public university in the United States and native speakers of Chinese were recruited from colleges in China. A pretest questionnaire designed by using Google Forms was used to collect background information from L2 learners, including self-rated proficiency in the four basic skills, age, gender, Chinese family background, years of learning Chinese, recent Chinese class, and other languages acquired. The L2 learners’ main biographical information and self-assessed proficiency scores were summarized in Table 1.

Participants (L2 learners and native speakers of Chinese) were individually assigned to the two main experiments (first acceptability judgment task and then self-paced reading task) and two-week delayed familiarity rating task. Native speakers of Chinese were not required to complete the familiarity rating task since these items, coming from the novice-level Chinese textbook, have been fully acquired by these adult native speakers during their preschool or elementary education.

### 3.2 Materials and procedure

#### 3.2.1 Determining Chinese FEs

Based on previous discussion, a Chinese FE can be defined as a linguistic category larger than a word, and it ranges from multi-word sequence (two or more than two

---

2 The self-reported proficiency score, along with three basic skills (speaking, listening and reading), was measured based on a scale from 1 to 10. For example, 1 indicates no proficiency, 3 indicates beginning or novice level, 5 indicates intermediate level, 7 indicates advanced level, and 10 indicates native-like proficiency. Here the reading score was separately included for it was highly associated with visual processing in this study.

---

### Table 1: Biographical information and proficiency scores for L2 participants (SDs in parentheses).

<table>
<thead>
<tr>
<th>L2 group (English as L1)</th>
<th>n</th>
<th>Age</th>
<th>Length (in years) of learning Chinese</th>
<th>Self-reported proficiency score (out of 10)</th>
<th>Self-reported reading score (out of 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning-intermediate</td>
<td>34</td>
<td>19.02 (1.02)</td>
<td>3.98 (4.15)</td>
<td>4.20 (0.75)</td>
<td>3.74 (0.74)</td>
</tr>
<tr>
<td>Advanced</td>
<td>29</td>
<td>26.18 (5.00)</td>
<td>8.36 (5.70)</td>
<td>7.11 (0.94)</td>
<td>7.25 (1.73)</td>
</tr>
</tbody>
</table>
words) to sentence pattern. All Chinese FEs were designed to be phrases consisting of a minimum of three characters or two words in this study. While FEs, such as collocations, include typical and atypical ones (Sonbul and Siyanova-Chanturia 2021), they were designed to be typical FEs for this study. In contrast, non-FE items were crafted to be atypical FEs, serving the purpose of comparison in experimental measurements.

Both subjective and objective methods were used to distinguish FEs from non-FE items for testing purposes, including L1 corpus frequency and the degree of formulaicity (or noncompositionality). An online survey (via Google Form) on ten Chinese native speakers’ (instructors of Chinese) intuition ratings was conducted to measure the degree of formulaicity using a 4-point Likert scale, where 1 represents “disagree”, 2 corresponds to “somewhat agree”, 3 stands for “agree”, and 4 denotes “strongly agree”. To decide formulaicity, these Chinese native speakers were asked to judge whether the test items (1) could be used as at least two words, and their components were more likely to combine together; and (2) could be treated as whole units (cannot be analyzed) in their language classrooms. To report, if eight out of ten judges rated either “3: agree” or “4: strongly agree” for a given test item, then it will be classified as an FE in this study. In this scenario, the remaining items (against the two properties) will be categorized as non-FE expressions. Due to the absence of a clearly defined threshold for distinguishing high and low frequency, relative frequency (log_{10} transformed) was employed to indicate that an FE’s frequency should be at least double that of a non-Fe’s, as observed in the 31 pairs of test items. For example, the frequency value of FE 太好了 is 4.32 while the non-Fe 太新了 is 1.53. The following criteria employed to create stimuli for the experiments were summarized in Table 2.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>FEs (typical)</th>
<th>Non-FE items (atypical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective measurement: intuition from Chinese language teachers (NSs)</td>
<td>(1) They can be used as at least two words, and their components are more likely to combine together.</td>
<td>Any ratings against the two properties of FEs</td>
</tr>
<tr>
<td></td>
<td>(2) All the Chinese teachers treat them as whole units (cannot be analyzed) in language classrooms.</td>
<td></td>
</tr>
<tr>
<td>Objective measurement: L1 corpus frequency (log_{10}-transformed; from BCC corpus(^a))</td>
<td>Relatively higher</td>
<td>Relatively lower</td>
</tr>
</tbody>
</table>

\(^a\)The link for BCC corpus is http://bcc.blcu.edu.cn/. The whole name for BCC corpus is BLCU Chinese Corpus, created by Endong Xun at Beijing Language and Culture University. The corpus is composed of 15 billion Chinese characters and covers a wide range of balanced databases with multiple genres, including newspapers, literature, Weibo posters, science and technology articles, comprehensive articles, and ancient Chinese texts.
3.2.2 Acceptability judgment task (AJT)

The AJT, implemented using PCIbex (Zehr and Schwarz 2018), assessed L2 learners’ real-time processing of FEs in Chinese. The task consisted of 124 items, with half intended to be acceptable and half unacceptable. Among the 62 acceptable items, 31 were FEs and 31 were non-FE items (See Appendix One in supplementary materials). To account for order effects, the order of test items was manipulated by balancing FE-first and non-FE-first presentations of test items. Participants were given written instructions and practiced with 5 items before proceeding to the 124 test items, which appeared randomly on the computer screen one at a time. The task involved reading and determining the acceptability of Chinese character strings. To move forward from one test item to another, participants needed to respond by pressing either the Z (YES) or M (NO) key on the keyboard, meanwhile, their response times (RTs) and judgment accuracy were recorded.

The acceptable items were generated as follows: an initial list of FE items was identified from the novice-level textbooks, *Integrated Chinese, Level 1 and 2* (Yao et al. 2016), used by L2 participants. Non-FE items were created by substituting one character in an FE item to form minimal pairs. Length was controlled, with FEs and non-FE items having the same mean number of characters (3.35). For example, the three-character FE 太好了 *tai hao le* ‘great’ in (1a) has a non-formulaic counterpart 太新了 *tai xin le* ‘too new’ in (1b). Although 太好了 is structurally similar to 太新了 (they both have the structure of “adverb 太 + adjective (好/新) + particle 了”), the meaning of 太新了 is more likely to be compositional rather than holistic based on the intuition judgment of Chinese NSs. Non-FE items were designed to be more flexible syntactically and semantically. Frequency was checked using the BCC corpus, and only those item pairs in which FEs exhibited frequencies at least twice as high as their corresponding non-FE items were retained. The final item list was determined in consultation with seven Chinese language instructors (2 males and 5 females) who

---

3 One of the reviewers argued that the component frequency of 好 is higher than 新, which might contribute to a processing advantage of 太好了 over 太新了. Ten pairs of FEs and non-FE items were purposefully selected to make sure the component frequency of FEs is lower than that of non-FEs. For instance, the FE 公共汽车 ‘bus’ and the non-FE item 英国汽车 ‘British car’ were chosen, as the component frequency of 公共 is lower than 英国 according to the BCC corpus. If there is still a processing advantage for these 10 pairs, as counterexamples, then the argument may not be valid. Statistical analysis revealed that these FEs still exhibited a processing advantage over non-FEs in the AJT but not in the SPR. This could be attributed to two possible factors: (1) the small size of the test sample, with only ten pairs of items as the dataset, and (2) participants’ lack of exclusive attention to the critical regions in context-based readings, resulting in no significant difference in RTs between FEs and non-FE items. However, when multicollinearity was considered in the model to analyze the complete 31 pairs of FEs and non-FE items, the result showed that FEs exhibited a significant processing advantage over non-FEs in both the AJT and SPR.
have taught these L2 participants using the aforementioned textbook. FE items were chosen based on high reported frequency and unanalyzability (more than 80%) in the survey on their classroom teaching. Idioms and other complex phrases, typically taught at advanced levels, were not included to ensure test suitability for learners of Chinese at different proficiency levels.

(1)  
a. 太好了  
*taɪ hɑːʊ le*  
‘great’ (FE)  
b. 太新了  
*taɪ xìŋ le*  
‘too new’ (non-FE)

The 62 unacceptable items were created either by changing the word order so that the string of characters no longer makes sense, or by including infelicitous collocations with verbs, classifiers, etc. The characters on the unacceptable items also came from the same Chinese textbooks and thus were assumed to be familiar to the learner participants. An example of each type was given in (2a) and (2b), respectively.

(2)  
a. *开会正在*  
*zhèngzǎi kǎihui*  
in process of having a meeting  
(Intended) ‘in process of having a meeting’  
b. *三只笔*  
*san zhī bǐ*  
three CL pen  
(Intended) ‘three pens’

### 3.2.3 Self-paced reading task (SPR)

The task was initially used in syntactic parsing studies (e.g., Ferreira and Henderson 1990; Mitchell 1987), it is an online computer-assisted technique where participants are asked to read sentences which have been segmented into words, in a line-by-line fashion they control by pressing a key to move forward (Marsden et al. 2018). RTs are recorded to measure participants’ sensitivity to specific linguistic phenomena (e.g., syntax and lexicon) in different contexts. The value of using SPR, compared with the acceptability judgment task, is to “allow researchers to determine with more precision the moment where difficulty (or processing cost) or facilitation (processing ease) arises, without seeking an explicit and offline judgment” (Marsden et al. 2018:
Inspired by Zheng et al. (2016), the study used a word-by-word reading paradigm, where each word is composed of one or more than one characters.

One hundred and twenty four stimuli, including 62 test sentences (See Appendix Two in supplementary materials; 31 target sentences with FEs and 31 baseline sentences with non-FE items, these FEs and non-FE items were from the AJT), and 62 distractor sentences, were randomized and presented to the participants through a self-paced reading task programmed with PCIbex. The sentences with FE-appear-first items and those without (non-FE-appear-first) were orderly balanced. This was done to prevent any potential impact of the initial presentation of FEs (or non-FE items) on the subsequent processing of their non-FE (or FE) counterparts during reading. Both FE and non-FE counterpart were embedded in the same sentence frame. There was one region of interest, which concerns the FE or non-FE counterpart. A yes-no question appeared after each sentence to ensure that the participants were reading for meaning and not just pressing the spacebar or key as fast as possible.

The non-cumulative moving-window method (McDonough and Trofimovich 2012) was adopted to present the self-paced reading. The target sentence was masked by dashes and spaces, with each segment revealed sequentially, covering up the previous segments. This allows tracking of reading time for each word or phrase and identifying processing difficulty. The task aims to measure participants’ reading time for critical displays (FEs) compared to control conditions (non-FE items). Sample design is shown below (“_” indicates a space):

**Target/Critical sentence** (an FE in the bracket indicates a critical region)

这_个_学_校_ [太好了], _他_认识了_很多_朋_友_。

‘This school is [great], he got to know many friends.’

**Control/Baseline sentence** (a non-FE counterpart in the bracket indicates a critical region)

这_个_学_校_ [太新了], _他_认识了_很多_朋_友_。

‘This school is [too new], he got to know many friends.’

**Question:** Did he know many friends in this school? (Y = Yes)

---

Note that in the SPR, the target sentences contain FEs and baseline sentences contain non-FE items, and both FEs and their non-FE counterparts were used in the previous AJT. The lists of AJT and SPR was pseudo randomized to reduce the priming effect of repeated use of the same test items. By doing so, the order of FEs and non-FE items was presented in totally different ways in the two experiments. To make sure its effectiveness, the design list (one list containing 62 critical items in AJT and another list including the same critical items in sentences (SPR)) was discussed with seven participants (three native speakers of Chinese and four L2 learners) in the pilot study. As a result, they could not detect an intended use of same test items in the two experiments, given the fact that the number of test items is large and the two experiments have different foci (AJT: process and judge each item without context; SPR: process each item in the context).
3.2.4 Familiarity rating task

The test was administered using Google Forms and conducted two weeks after the first two experiments. There are only critical items in the rating task, including all 31 FEs and 31 non-FE items, which were used in both AJT and SPR. The rating was designed on a scale of 1–5, including “1 = Not at all familiar, 2 = Slightly familiar, 3 = Moderately familiar, 4 = Very familiar, 5 = Extremely familiar”. The goal is to collect participants’ ratings on the degree of familiarity with the critical items. Meanwhile, participants were required to translate all these items into English only when they were familiar with them. The goal is to determine how participants handle these items (through morphosyntactic analysis or holistic processing) and assess their accuracy in translating them to demonstrate genuine familiarity. The translations provided by the participants were finally evaluated by four bilingual speakers proficient in both Chinese and English to ensure inter-rater reliability. The task, with a newly-organized order of item presentation, was set apart from previous experiments in order to minimize the priming effect on familiarity ratings.

3.3 Data analysis

Statistical analyses were conducted in R (R Core Team 2021). Two sets of mixed-effects models were fitted: one for data from all participants and another for data from learners only. For models including all participants, the fixed predictors were proficiency and condition. Proficiency had three levels: beginning-intermediate, advanced, and native. Helmert coding was applied, comparing NS and NNS first, followed by a comparison of beginning-intermediate and advanced learners. Condition had two levels: FEs and non-FE items. Sum coding was applied. For models including only learner participants, proficiency, condition, and familiarity were entered as predictors. Proficiency had two levels: beginning-intermediate, advanced, and sum coding was applied. Familiarity, measured on a scale of 1–5, was treated as a continuous variable in the analyses. Familiarity was included only for the learner models because no familiarity data was collected from Chinese NSs. The mixed-effects models incorporated maximal random effects structure whenever possible (Barr et al. 2013), including participants and test items as random effects. Initially, all random intercepts and random slopes were included, and then the random structure was gradually simplified using maximum likelihood estimation. The variance inflation factor (VIF) values for each predictor (proficiency, condition, familiarity) were examined to control for multicollinearity. As Kapteijns and Hintz (2021) suggested, predictors with VIF values surpassing 5 are considered problematic. The VIF
for the fitted model was thus checked and the fixed effects structures were simplified until the VIF fell below the standard threshold of 5.0.

The data from AJT were analyzed with mixed-effects regression models using the lme4 package (Bates et al. 2015) in R. Separate models were fitted for accuracy and RT. For the RT models, the following steps were followed for data preprocessing. First, the data excluded incorrect judgments (1,194 responses out of total 5,828 responses from NSs and NNSs) on critical items. Second, responses that took shorter than 100 ms were also excluded. Third, RTs that fell outside 2.5 standard deviations from the average for each participant were also removed. Such cutoffs were adopted after visually examining exploratory graphics illustrating the distribution of subjects’ RTs. Data loss for the second and third steps was 2.09 %.

The RTs on the critical regions and spill-over regions of the SPR task were analyzed. No spillover effects were observed, so only the RTs on the critical regions were reported. For data preprocessing, first, responses that took shorter than 100 ms to the critical regions were excluded. Then, RTs that fell outside 2.5 standard deviations from the average for each participant were removed. These cutoffs were determined by visually analyzing exploratory graphics that displayed the distribution of subjects’ RTs, in accordance with the data cleaning methods for SPR used by Jegerski (2014) and Shi et al. (2023). Data loss was 2.62 % as only one participant’s data was removed in the second step.

4 Results

The findings are presented in alignment with the research questions. The answers to RQ1 were derived from the RT and accuracy data of the AJT (without context) and the RT data of the SPR (with context). For RQ2, the answers were obtained by examining the familiarity and the proficiency factor within the statistical models of AJT and SPR. Answers to RQ3 were found in the RT and accuracy data of NSs and NNSs across both models.

4.1 Familiarity rating task

Table 3 presents the mean familiarity ratings of the L2 learners on the critical items. L2 participants showed higher familiarity with FEs compared to non-FE items (FEs: $M = 4.5, SD = 1.07$; Non-FEs: $M = 3.5, SD = 1.41$). It was also revealed that advanced learners rated both FEs and non-FE items higher than beginning-intermediate learners (Beginning-intermediate: FEs: $M = 4.3, SD = 1.28$; Non-FEs: $M = 3.3, SD = 1.46$; Advanced: FEs: $M = 4.7, SD = 0.86$; Non-FEs: $M = 3.7, SD = 1.37$).
4.2 Acceptability judgment task

Descriptive statistics of mean RTs in the AJT (Figure 1) showed that FEs generally took shorter to respond ($M = 3,383.3$, $SD = 1,124.6$) than non-FE items ($M = 4,486.2$, $SD = 1,512.3$). Moreover, there was a general trend that as proficiency level increased, there was a consistent decline in RTs between FEs and non-FE items for all groups. The gap of RTs between FEs and non-FE items for each learner’s group was greater than that of the NS group. The results for the mixed effects model on RT for all participants were presented in Table 4. The main effect of the condition was significant ($estimate = -0.13$, $SE = 0.01$, $t = -13.64$, $p < 0.001$), showing that all participant groups processed FEs differently from non-FE items in the AJT. The main effect of the proficiency (NSs, NNSs) ($estimate = -0.78$, $SE = 0.07$, $t = -11.10$, $p < 0.001$) also played a significant role in RTs, suggesting a difference in processing time between NSs and NNSs in the AJT. There was a significant two-way interaction between condition and proficiency (NSs, NNSs) ($estimate = 0.09$, $SE = 0.02$, $t = 4.55$, $p < 0.001$), indicating that the effects of the condition varied between native and learner groups. However, there was no interaction between condition and proficiency (beginning-intermediate vs. advanced) ($estimate = 0.01$, $SE = 0.02$, $t = 0.51$, $p > 0.05$), indicating there was no clear difference between beginning-intermediate and advanced learners when they processed FEs and non-FE items. Pairwise comparisons based on proficiency and condition were then conducted to understand these interaction effects. The analysis revealed that while there was a processing advantage for FEs over non-FE items among all groups (NSs: $estimate = -0.13$, $SE = 0.02$, $t = -4.26$, $p < 0.001$; Advanced learners: $estimate = -0.31$, $SE = 0.03$, $t = -8.89$, $p < 0.001$; Beginning-intermediate learners: $estimate = -0.33$, $SE = 0.03$, $t = -10.50$, $p < 0.001$), the effects were more pronounced among L2 learners than among native speakers.

In the learner model, Table 5 showed a significant main effect of condition ($estimate = -0.13$, $SE = 0.02$, $t = -5.65$, $p < 0.001$) on RTs, suggesting learners’ processing of FEs differed from that of non-FE items in the AJT. Familiarity also showed a significant main effect ($estimate = -0.05$, $SE = 0.01$, $t = -6.04$, $p < 0.001$) in the AJT, revealing that learners’ processing time was influenced by the degree of familiarity

<table>
<thead>
<tr>
<th>Group</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FEs</td>
</tr>
<tr>
<td>Beginning-intermediate L2</td>
<td>4.3 (1.28)</td>
</tr>
<tr>
<td>Advanced L2</td>
<td>4.7 (0.86)</td>
</tr>
</tbody>
</table>
with the critical items. However, proficiency did not have a significant main effect 
\(\text{estimate} = 0.06, \ SE = 0.05, \ t = 1.06, \ p > 0.05\) on learners’ RTs, manifesting no sub-
stantial difference in RTs between beginning-intermediate learners and advanced
learners. There was a significant two-way interaction between familiarity and pro-
\(\text{ficiency} = -0.02, \ SE = 0.01, \ t = -3.30, \ p < 0.001\), revealing that familiarity
played varying roles in learners’ processing across different levels. This was ob-
erved in Figure 2, where beginning-intermediate learners were less affected by

\[\text{Table 4: Estimates of fixed effects comparing NNSs and NSs in RTs for FEs and non-FE items in AJT.}\]

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>df</th>
<th>t-Value</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8.09</td>
<td>0.03</td>
<td>86.56</td>
<td>228.26</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Condition (FEs, non-FEs)</td>
<td>-0.13</td>
<td>0.01</td>
<td>62.90</td>
<td>-13.64</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (NSs, NNSs)</td>
<td>-0.78</td>
<td>0.07</td>
<td>69.26</td>
<td>-11.10</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (NSs, NNSs) \times condition</td>
<td>0.09</td>
<td>0.02</td>
<td>58.31</td>
<td>4.55</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (beginning-intermediate, advanced)</td>
<td>-0.18</td>
<td>0.08</td>
<td>70.19</td>
<td>-2.35</td>
<td>&lt;0.05*</td>
</tr>
<tr>
<td>Proficiency (beginning-intermediate, advanced) \times condition</td>
<td>0.01</td>
<td>0.02</td>
<td>67.49</td>
<td>0.51</td>
<td>&gt;0.05</td>
</tr>
</tbody>
</table>

\[\text{Note.} \ *p < 0.05, \ **p < 0.01, \ ***p < 0.001.}\]
familiarity, compared with advanced learners. Specifically, advanced learners dominantly assigned high familiarity (4–5 rating) to both FEs (green dots) and non-FE items (red dots) while there were more balanced low ratings (1–3) from beginning-intermediate learners. It was also evident that familiarity might not be the sole factor affecting the item processing among learners at two levels. As the majority of data points in Figure 2 clustered towards the lower end of higher familiarity ratings, specifically around ratings 4–5, FEs (green dots) still had generally shorter RTs than non-FE items (red dots) for learners of Chinese at two levels.

Regarding accuracy in the AJT, descriptive statistics in Figure 3 depicted a trend that the judgments of FEs had much higher accuracy (M = 92.2 %) than non-FE items (M = 66.8 %). Moreover, the judgment accuracy from FEs to non-FE items consistently

Table 5: Estimates of fixed effects comparing two levels of NNSs in RTs for FEs and non-FE items in AJT.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>df</th>
<th>t-Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8.62</td>
<td>0.06</td>
<td>166.71</td>
<td>132.78</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Condition (FEs, non-FEs)</td>
<td>−0.13</td>
<td>0.02</td>
<td>48.38</td>
<td>−5.65</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (beginning-intermediate, advanced)</td>
<td>0.06</td>
<td>0.05</td>
<td>201.74</td>
<td>1.06</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td>Familiarity</td>
<td>−0.05</td>
<td>0.01</td>
<td>2,013.67</td>
<td>−6.04</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency × familiarity</td>
<td>−0.02</td>
<td>0.01</td>
<td>1,812.91</td>
<td>−3.30</td>
<td>&lt;0.001***</td>
</tr>
</tbody>
</table>

Note. *p < 0.05, **p < 0.01, ***p < 0.001.

Figure 2: Individual RTs (in milliseconds) of FEs and non-FE items by L2 group, grouped by item familiarity in AJT (Note: the slope of a line is the change in RT over the change in familiarity).
declined across all different groups. The logistic mixed effects model (Table 6) included data from NSs and NNSs and it revealed a significant main effect of condition ($estimate = 1.11, SE = 0.07, z = 14.72, p < 0.001$), indicating that FEs received significantly different (higher) accuracy rates than non-FE items from NSs in the AJT. There was also a significant main effect of proficiency (NSs vs. NNSs) ($estimate = 2.13, SE = 0.22, z = 9.33, p < 0.001$), showing that proficiency significantly affected their judgment accuracy. As Figure 3 showed, the judgment accuracy for FEs from NSs ($M = 99.5\%$) was noticeably higher than among NNSs ($M = 88.55\%$) in the AJT. Generally, significant differences in accuracy between FEs and non-FE items were observed among NSs, indicating that NSs made more accurate judgments on FEs.

The L2 learner model for AJT accuracy was presented in Table 7. The accuracy data for the learners’ groups revealed that familiarity had a significant main effect ($estimate = 0.82, SE = 0.05, z = 16.74, p < 0.001$) on accuracy, indicating that their

---

**Table 6:** Estimates of fixed effects comparing NSs and NNSs in accuracy for FEs and non-FE items in AJT.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$z$-Value</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.20</td>
<td>0.17</td>
<td>12.95</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Condition (FEs, non-FEs)</td>
<td>1.11</td>
<td>0.07</td>
<td>14.72</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (NSs, NNSs)</td>
<td>2.13</td>
<td>0.22</td>
<td>9.33</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (beginning-intermediate, advanced)</td>
<td>0.85</td>
<td>0.23</td>
<td>3.69</td>
<td>&lt;0.001***</td>
</tr>
</tbody>
</table>

*Note. *$^a p < 0.05, ^b p < 0.01, ^c p < 0.001$. 

---

**Figure 3:** Mean accuracy (in percentage) of FEs and non-FE items in AJT.
judgment accuracy was influenced by the degree of familiarity with the critical items. Also, condition played a significant role (estimate = 0.88, SE = 0.12, z = 7.43, p < 0.001) in accuracy, indicating FEs differed from non-FE items in affecting learners’ accuracy in the AJT. However, proficiency had no significant impact (estimate = −0.06, SE = 0.19, z = −0.35, p > 0.05) on accuracy rates, indicating no significant difference between beginning-intermediate learners and advanced learners. There was a significant interaction between familiarity and proficiency (estimate = 0.11, SE = 0.04, z = 2.54, p < 0.05), showing learners at two proficiency levels had different familiarity with critical items. Figure 4 depicted the influence of familiarity and

Table 7: Estimates of fixed effects comparing two levels of NNSs in accuracy for FEs and non-FE items in AJT.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>z-Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−1.58</td>
<td>0.25</td>
<td>−6.27</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Condition (FEs, non-FEs)</td>
<td>0.88</td>
<td>0.12</td>
<td>7.43</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (beginning-intermediate, advanced)</td>
<td>−0.06</td>
<td>0.19</td>
<td>−0.35</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td>Familiarity</td>
<td>0.82</td>
<td>0.05</td>
<td>16.74</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency × familiarity</td>
<td>0.11</td>
<td>0.04</td>
<td>2.54</td>
<td>&lt;0.05*</td>
</tr>
</tbody>
</table>

Note. *p < 0.05, **p < 0.01, ***p < 0.001.

Figure 4: Individual accuracy (in percentage) of FEs and non-FE items by L2 group, grouped by item familiarity in AJT (Note: the slope of a line is the change in accuracy over the change in familiarity; “0” marks incorrect judgment while “1” marks correct judgment).
proficiency on processing accuracy, showing that compared with advanced learners, beginning-intermediate learners were less influenced by the effect of familiarity in their judgment accuracy. This was evident in the pattern where the beginning-intermediate learner group exhibited a more even distribution of dots across the low familiarity rating region (1–3), in contrast to the same region in the advanced learner group. In sum, L2 learners across two proficiency levels demonstrated greater sensitivity to FEs compared to non-FE items, as reflected in their higher accuracy rates.

4.3 Self-paced reading task

Descriptive statistics (Figure 5) revealed a consistent decrease in RTs for the critical regions in the SPR, as the proficiency levels of speakers advanced. Among learners at both levels, FEs had, on average, shorter RTs \( (M = 1731.35, SD = 742.07) \) compared to non-FE items \( (M = 2,505.7, SD = 471.25) \). However, native speakers showed similar RTs for both FEs \( (M = 358.7, SD = 148.43) \) and non-FE items \( (M = 368.5, SD = 107.29) \). Overall, learners of Chinese read FEs faster than non-FE items, whereas NSs read the two item types similarly in terms of RTs in the SPR.5

Table 8 revealed a significant main effect of condition \( (estimate = -0.10, SE = 0.01, t = -11.52, p < 0.001) \) in the SPR, indicating all groups of participants processed FEs markedly differently compared with non-FE items. The main effect of proficiency (NSs vs. NNSs) was also significant \( (estimate = -1.60, SE = 0.11, t = -14.22, p < 0.001) \), suggesting that processing time varied notably between the NS group and the NNS groups. There was a significant interaction between condition and proficiency (NSs vs. NNSs) \( (estimate = 0.15, SE = 0.02, t = 7.51, p < 0.001) \), showing a notable difference in RTs between FEs and non-FE items for both NSs and NNSs. Pairwise comparisons were conducted to reveal that there was no significant difference \( (estimate = -0.01, SE = 0.03, t = -0.40, p > 0.05) \) for NSs to process FEs and non-FE items.

In learners’ data (Table 9), the main effect of the condition was significant \( (estimate = -0.12, SE = 0.02, z = -5.11, p < 0.001) \), indicating NNSs’ sensitivity to FEs and non-FE items in terms of the processing time. There was also a significant main effect of familiarity \( (estimate = -0.06, SE = 0.01, z = -7.69, p < 0.001) \) on learners’ RTs,

---

5 One of the reviewers mentioned the possibility of a spillover effect in the SPR, suggesting that a slowdown or acceleration in reading might occur not precisely at the critical region but at the subsequent word. Although a significant main effect was identified for the condition (critical regions) on RTs in the SPR, as detailed in Tables 8 and 9, it was essential to study whether a similar effect is evident in the spillover (post-critical) regions for both L2 and NS groups. However, the results indicated no significant effect of spillover regions on RTs, affirming that the main differences in RTs between FEs and non-FE items were attributed to the critical regions.
suggesting the degree of familiarity with critical regions (FEs and non-FE items) played a role in learners’ processing speed. Proficiency did not exert a significant main effect (estimate = 0.02, SE = 0.06, z = 0.36, p > 0.05) on learners’ processing time. The interaction between familiarity and proficiency was significant (estimate = −0.02, SE = 0.01, t = −3.09, p < 0.01), however, indicating learners’ RTs for critical regions were influenced by familiarity, and this influence varied across two proficiency levels in the SPR. As depicted in Figure 6, there was a difference in the

Figure 5: Mean RT (in milliseconds) of FEs and non-FE items by participant group in SPR. Error bars represent standard error of the mean.

Table 8: Estimates of fixed effects comparing NSs and NNSs in RTs for FEs and non-FE items in SPR.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>df</th>
<th>t-Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.85</td>
<td>0.05</td>
<td>91.82</td>
<td>119.77</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Condition (FEs, non-FEs)</td>
<td>−0.10</td>
<td>0.01</td>
<td>70.92</td>
<td>−11.52</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (NSs, NNSs)</td>
<td>−1.60</td>
<td>0.11</td>
<td>70.98</td>
<td>−14.22</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (beginning-intermediate, advanced)</td>
<td>−0.23</td>
<td>0.12</td>
<td>70.98</td>
<td>−1.88</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td>Condition × proficiency (NSs, NNSs)</td>
<td>0.15</td>
<td>0.02</td>
<td>70.82</td>
<td>7.51</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Condition × proficiency (beginning-intermediate, advanced)</td>
<td>−0.01</td>
<td>0.02</td>
<td>71.00</td>
<td>−0.13</td>
<td>&gt;0.05</td>
</tr>
</tbody>
</table>

Note. *p < 0.05, **p < 0.01, ***p < 0.001.
impact of familiarity on RTs between beginning-intermediate and advanced level groups. Similar to the scatter plot in the AJT, beginning-intermediate learners in the SPR showed less influence from familiarity on their RTs compared to advanced learners. It became clear that the advantage of processing FEs in L2 learners might not be entirely due to familiarity. Figure 6 showed a greater concentration of green dots (FEs) than red ones (non-FEs) around the bottom of the high familiarity rating 5. This demonstrated that FEs were still processed faster than non-FE items even if familiarity was controlled to the same extent.

Table 9: Estimates of fixed effects comparing two levels of NNSs in RTs for FEs and non-FE items in SPR.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>df</th>
<th>t-Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.65</td>
<td>0.08</td>
<td>90.78</td>
<td>92.46</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Condition (FEs, non-FEs)</td>
<td>−0.12</td>
<td>0.02</td>
<td>42.33</td>
<td>−5.11</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency (beginning-intermediate, advanced)</td>
<td>0.02</td>
<td>0.06</td>
<td>134.22</td>
<td>0.36</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td>Familiarity</td>
<td>−0.06</td>
<td>0.01</td>
<td>2,457.87</td>
<td>−7.69</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Proficiency × familiarity</td>
<td>−0.02</td>
<td>0.01</td>
<td>1,678.17</td>
<td>−3.09</td>
<td>&lt;0.01**</td>
</tr>
</tbody>
</table>

Note. *p < 0.05, **p < 0.01, ***p < 0.001.
5 Discussion

The ensuing discussion section offers further explanations for the data analysis. The discussion of RQ1 can be found in Section 5.1. Sections 5.2 and 5.3 focus on interpreting the answers to RQ2. The discussion of RQ3 is included in Sections 5.1 and 5.3. Section 5.4 summarizes the limitations and future directions.

5.1 The advantage of holistic processing: being fast with fewer errors

The results from the two online experiments (AJT and SPR) consistently showed that FEs (low compositionality in this study) had a processing advantage compared with non-FE items in learners’ groups. In the AJT, test items concerning FEs were read faster with fewer errors (higher accuracy) than their corresponding non-formulaic parts. In SPR, we see the same pattern for RT in the processing of FEs by learners at the two levels. While the advantages of FEs in terms of rapid processing may not directly substantiate the Holistic Hypothesis, as exemplified by Siyanova-Chanturia (2015), the high level of processing accuracy serves to confirm the influence of holistic storage in online processing. As evident from Figure 3, L2 learners consistently display a significantly higher level of accuracy in processing FEs when compared to non-FE items. This underscores the notion that the ease with which L2 learners access FEs, achieving quicker processing with greater accuracy, is likely attributed to the well-established storage of FEs in their learning experiences. Because of the entrenched storage of FEs, L2 learners did not have to conduct a character-by-character analysis; instead, they retrieved these patterns as holistic or single units at one time from their memory, which suggests a fast-and-accurate reading process. The findings are coherent with the usage-based statement that multiword sequences such as FEs are represented as single units, being stably stored and directly retrieved from the mental lexicon as wholes, especially when learners receive frequent exposure in language acquisition (e.g., Ellis 2003; Jiang and Nekrasova 2007; Schmitt et al. 2004; Underwood et al. 2004). The subjective selection of FEs in this study was guided by instructors who treated them as frequently used and unanalyzable structures in classroom learning, which could aid learners in perceiving, recognizing, and storing these FEs as holistic units in their long-term memory. In addition, L2 learners across two proficiency levels both had the processing advantage for FEs in the AJT and SPR, whereas adult Chinese NSs only showed the processing advantage in the AJT. It is understandable that compared with the sentence reading task, both NSs and NNSs might take more time to examine each item in the judgment task,
allowing them to focus more on the semantics and syntax of each item. The conventionalized structure of FEs may make participants feel easier to process compared to non-FE items, based on their intuition and learning experiences. However, in the SPR, NSs might confidently read both FEs and non-FE items quickly in order to progress until they finish reading the entire sentences and can answer the questions. Therefore, it seems that NSs’ ability to read critical regions did not significantly differ from their reading of other non-critical regions when context was included. Additionally, considering that all the characters used originates from the novice-level Chinese textbook, it is anticipated that these NSs should possess ample confidence and fluency to comprehend all critical items equally well. The result provides evidence from L2 Chinese to show that the holistic processing of FEs can be more commonly found in NNSs than NSs, which differs from previous observations that indicated a processing advantage for L2 learners only (e.g., Conklin and Schmitt 2012; Jiang and Nekrasova 2007; Wolter and Yamashita 2018).

Another possible finding shows that the processing advantage of FEs seems to be more context-dependent for L2 learners, although condition played a significant role in both the AJT and SPR. In the context-free AJT, the RTs for FEs were higher for beginning-intermediate learners (4,414 ms) compared to advanced learners (3,853 ms). In contrast, SPR results revealed lower RTs in processing FEs, with 1,895 ms for beginning-intermediate learners and 1,567 ms for advanced learners. Compared with the AJT, learners as participants spent less time in the SPR where contexts were provided in different sentences. However, we should notice that RT in the AJT task can reflect both processing time and judgment time, whereas RT in the SPR likely only reflects processing time. For instance, L2 participants might spend more time thinking and pressing the key to judge each critical test item as a separate question in the AJT. In the SPR, however, they might more quickly read each critical test item by unremittingly pressing the key to finish the sentence reading. It is possible that the processing advantage of FEs could be affected by the context, as participants might more easily and quickly process FEs with the help of preceding information in sentence reading. We know the fact that Chinese NSs did not have the processing advantage for FEs when context was included, whereas the processing advantage for FEs was still pronounced among L2 learners in the context-based reading. This suggests that context might have a trade-off effect on the processing advantage for NSs rather than for L2 learners of Chinese. Possible reasons could include participants’ reading fluency and habits, as discussed previously. However, it remains unclear how exactly and to what extent context can influence their online processing of FEs, which is worth exploring in future research. The finding is nevertheless in concordance with the findings from previous studies, such as Zheng et al. (2016), where they employed multiple regression
models to conclude that the effect of context in SPR can significantly reduce the processing advantage of FEs, and this effect exists among intermediate-level L2 learners as their participants.

5.2 The effects of familiarity

It has been acknowledged that input frequency, including the exposure to the word-level components and phrase-level structures, can play a driving role in facilitating L2 learners’ on-line processing of collocations (a type of FEs) in context-free experiments (e.g., Wolter and Gyllstad 2013; Wolter and Yamashita 2018). However, the token frequency of test items might not be an accurate indicator of L2 learners’ processing advantage in real-time experiments. Instead, this study chose to measure item familiarity as another crucial factor, which can be more directly associated with the instantaneous recognizing and reading of FEs and non-FE items in L2 learners’ processing. The significant main effect of familiarity on RTs and accuracy in the AJT, along with the significant main effect of familiarity on RTs in the SPR, suggests that familiarity enhances both the speed and correctness of visual processing in L2 learners. These findings align with previous research on the role of familiarity (e.g., Carrol and Conklin 2017; Carrol et al. 2016) in L2 processing of FEs.

However, what sets it apart is that the advantage was not biased towards FEs, as there were numerous overlaps between the data points of FEs and non-FE items in Figures 4 and 6. This indicates a general tendency in lexical or grammatical tests, where participants’ responses are facilitated as they become familiar with test items, regardless of item types. Although previous analyses have emphasized the importance of familiarity in promoting the processing of FEs and non-FE items, we should be noted that the processing advantage cannot be attributed to the familiarity effect alone, given that the holistic nature (formulaicity or non-compositionality) of FEs can also trigger quicker RTs compared with non-FE items. As mentioned in the scatter plots for the AJT and SPR, even for the critical items with which the learners of Chinese were extremely familiar (rated as five in the familiarity rating task), most learners at the same level still took less time to respond to FEs than what they did for non-FE items. When controlling for familiarity at the same level, the processing advantage associated with FEs persisted. This indicates that the processing advantage might also be provided by the inherent property (holistic status) of FEs within L2 learners’ mental lexicon. However, future research will examine whether familiarity has a greater impact on enhancing processing advantages than the condition effect does.
5.3 The effect of proficiency

Despite the consistently decreasing trend in RTs from NNS groups to the NS group, proficiency level did not exert a significant influence on the context-free and context-dependent processing of FEs and non-FE items within NNS groups. Chinese NSs, although exhibiting similar RTs in processing FEs and non-FE items, required significantly less time on average to read the test items compared to L2 learners. This is not surprising, given the fact that the higher participants’ proficiency levels are, the more input and retention they can acquire, making it easier and faster for them to recognize and comprehend the lexical items (with increased familiarity as well). The different performances between NSs and NNSs of Chinese on the two experiments speak to Wray’s (2002, 2008) claims that natives and L2 learners process FEs in fundamentally different ways. As discussed before, there was no processing advantage of FEs in terms of NSs’ RTs in the SPR. The processing advantage of FEs was more pronounced in the learners’ group than in the NSs’ group in the AJT. These observations also suggest that the processing advantage of FEs might be counterbalanced or neutralized when participants’ proficiencies approach the native-like levels. This trend is also evident in both scatter plots in the AJT (Figure 4) and SPR (Figure 6), where the RT gap between FEs (green lines) and non-FE items (red lines) showed a tendency to slightly decrease from low-level to high-level learners, despite the absence of a significant proficiency effect between the two groups. The finding generally contradicts Ellis and Simpson-Vlach’s (2009) argument that NSs tend to be sensitive from their long usage histories to FEs so that they preferentially process them. Instead, the data analysis suggests that there is no outstanding difference for NSs in processing Chinese FEs and non-FE items, especially in context-based reading, due to their fully-acquired L1 knowledge in their minds. Unexpectedly, there was a significant deficiency in the influence of proficiency levels on the performances of L2 learners, in the two experiments. A potential explanation may lie in the utilization of characters derived from basic-level Chinese textbooks, ensuring test accessibility for all L2 learners. However, this choice has led to a limited difference in RTs between low-level and high-level learners (no significant effect of proficiency), as they were universally exposed to these foundational characters early in their learning journeys. Another plausible account is the lack of L2 learners’ study abroad experience, which contributed to minimizing distinctions in online performance among individuals assigned to varying proficiency levels. The result thus does not seem to support the positive correlation between proficiency and the processing advantage in NNS groups, as demonstrated in previous studies (e.g., Conklin and Schmitt 2012; Nekrasova 2009; Wolter and Yamashita 2018).
5.4 Limitation of this study

The study has limitations as it did not account for participants’ sensitivity to congruency between their L1 and L2 due to the constraints imposed by the limited number of characters available for the experimental design. We know that the similarity between learners’ L1 and L2 knowledge can enhance FE processing (e.g., Chen 2024; Ding and Reynolds 2019; Wolter and Gyllstad 2011, 2013; Yamashita and Jiang 2010). A bilingual model of language processing (Conklin and Carrol 2018) could explain the facilitative effect of Chinese FEs with natural English translations. In the present study, all the critical items have English equivalents, owing to the semantic and pragmatic similarities in Chinese and English words. Since the congruency effect likely influences both FEs and non-FE items, it is reasonable to assume that the L1 transfer would not confound the explanation for the processing advantage of FEs. In future studies, it is important to expand the range of Chinese lexical items and include congruent and incongruent items to thoroughly measure the congruency effect. It is also essential to augment the sample size with more stratification (e.g., learners with three levels: novice, intermediate, advanced) to examine if and how processing advantage can significantly interact with proficiency levels.

6 Conclusion

This study demonstrated the Holistic Hypothesis by mainly measuring how L2 learners of Chinese processed FEs and non-FE items using the two experimental paradigms (AJT and SPR).

The findings can be summarized based on the research questions: (1) L2 learners at each proficiency level demonstrated a processing advantage for FEs over non-FE items in both context-free (AJT) and context-dependent (SPR) tasks; (2) Familiarity and holistic nature both influenced learners’ processing of FEs, while the proficiency effect only manifested when comparing NS with NNS groups; (3) The processing of FEs differed between NSs and NNSs. Learners of Chinese showed faster processing of FEs compared to non-FE items in reading, regardless of context. However, NSs only showed a similar pattern in context-free reading (AJT).

The findings extended the advantage of holistic processing to L2 Chinese by empirically examining FEs as a type of conventional linguistic pattern. The attested processing advantage in FEs brings insights into the connection between usage-based theory and chunking in the cognitive approach to L2 learning, shaping our understanding of how learners of Chinese comprehend and recognize FEs as visual stimuli and how they process FEs in different experimental settings.
Acknowledgments: I wish to express my heartfelt thanks to Dr. Janet Nicol and other dissertation committee members for their constructive suggestions on the early version. Additionally, I am grateful to Jue Wang for assisting with the statistical analysis. I also deeply appreciate the feedback from the reviewers and the data contribution from research participants.

References


Arnon, Inbal & Eve V. Clark. 2011. Why brush your teeth is better than teeth: Children’s word production is facilitated in familiar sentence-frames. Language Learning and Development 7(2). 107–129.


Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/iral-2023-0262).