Applied Psycholinguistics (2022), **43**, 363–388
doi:10.1017/S0142716421000552

ORIGINAL ARTICLE

NS and NNS processing of idioms and nonidiom formulaic sequences: What can reaction times and think-alouds tell us?

Hang Zheng[1]*, Melissa A. Bowles[2] and Jerome L. Packard[3]

[1]Department of Chinese (Zhuhai), Sun Yat-Sen University, Zhuhai, China, [2]Department of Spanish and Portuguese, University of Illinois at Urbana-Champaign, Urbana, USA and [3]Department of East Asian Languages and Cultures, University of Illinois at Urbana-Champaign, Urbana, USA
*Corresponding author. Email: zhengh73@mail.sysu.edu.cn

(Received 10 September 2020; revised 15 September 2021; accepted 26 October 2021; first published online 15 December 2021)

Abstract

Although researchers generally agree that native speakers (NSs) process formulaic sequences (FSs) holistically to some extent, findings about nonnative speakers (NNSs) are conflicting, potentially because not all FSs are psychologically equal or because in some studies NNSs may not have fully understood the FSs. We address these issues by investigating Chinese NSs and NNSs processing of idioms and matched nonidiom FSs in phrase acceptability judgment tasks with and without think-alouds (TAs). Reaction times show that NSs processed idioms faster than nonidioms regardless of length, but NNSs processed 3-character FSs faster than 4-character FSs regardless of type. TAs show NSs' understanding of FSs has reached ceiling, but NNSs' understanding was incomplete, with idioms being understood more poorly than nonidioms. Although we conclude that idioms and nonidioms have different mental statuses in NSs' lexicons, it is inconclusive how they are represented by NNSs. TAs also show that NNSs employed various strategies to compensate for limited idiom knowledge, causing comparable processing speed for idioms and nonidioms. The findings highlight the importance of distinguishing subtypes of FSs and considering NNSs' quality of understanding in discussions of the psychological reality of FSs.

Keywords: formulaic sequences; idioms; reaction time; think-alouds; quality of understanding

Formulaic sequences (FSs) are prefabricated word bundles that are recurrent in language and possess highly conventional meanings. As mounting evidence has shown that the acquisition of FSs is closely related to overall language proficiency (Boers et al., 2006; Dai & Ding, 2010; Lindstromberg & Boers, 2008; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Weinert, 1995), increasing attention has been given to how native speakers (NSs) and nonnative speakers (NNSs) process FSs. However, as Myles and Cordier (2017) pointed out, many studies gauging the processing of FS are unclear in defining what type of FS they are investigating and often generalize the processing advantage found for a single type of FS to all formulaic language, claiming that FSs are holistically stored in long-term memory. Nevertheless, the fact that FSs are processed faster does not necessarily indicate that they are represented as a whole in the mental lexicon (Siyanova-Chanturia, 2015). In addition, the great diversity in types of FSs makes it untenable to treat all FSs as if they were psychologically the same (Boers & Lindstromberg, 2012). Recent research (Carrol & Conklin, 2020) has found that NSs' processing of different types of FSs is regulated by different linguistic properties (i.e., idioms by frequency, familiarity, and decomposability; binominals by predictability and semantic association; collocations by mutual information). This complication in turn poses difficulties for the integration of findings about FS processing into FS acquisition because NSs and NNSs do not perceive and use all FSs in the same way (Nekrasova, 2009). The present study set out to address the issue by comparing the online processing of two subtypes of FSs and investigating the extent to which processing speed by NSs and NNSs is related to how well they understand FSs.

In the FS literature, a sizable number of studies have employed online instruments, such as reaction/reading times or eye-tracking paradigms, to investigate NSs and/or NNSs' processing patterns (e.g., Carrol & Conklin, 2020; Conklin & Schmitt, 2008; Gyllstad & Wolter, 2016; Jiang & Nekrasova, 2007; Jiang et al., 2020; Siyanova-Chanturia et al., 2011; Tabossi et al., 2009; Tremblay et al., 2011; Underwood et al., 2004; Vilkaitė & Schmitt, 2019; Wolter & Yamashita, 2014, 2018; Yamashita & Jiang, 2010). Many studies have utilized offline measures, such as metalinguistic ratings, controlled production, or verbal reports, to assess the status of speakers' FS knowledge (e.g., Bardovi-Harlig & Stringer, 2017; Cieślicka, 2006; Cooper, 1999; Irujo, 1986; Kim, 2016; Martinez & Murphy, 2011; Nekrasova, 2009; Siyanova-Chanturia & Janssen, 2018; Spöttl & McCarthy, 2004; Van Lancker Sidtis, 2003). The present study goes one step further by triangulating two concurrent measures, reaction time (RT) and think-aloud (TA) protocols, to examine how NSs and NNSs process two different types of FSs, namely idioms and matched non-idiom FSs, to determine if there is a relationship between how fast speakers respond to and how well they understand the two types of FSs.

Literature review

There are two main strands of FS research: processing-based research and comprehension-based research. The former focuses on the holistic nature and representation of FSs in the mental lexicon, while the latter examines how well language users know FSs and what strategies they employ to comprehend FSs. Most studies test only NSs, or only NNSs. Only a few have juxtaposed NSs and NNSs, comparing the psychological status and/or acquired knowledge about FSs in the first and second languages (L1 and L2). These studies are reviewed in detail below.

Underwood et al. (2004) investigated how NSs and NNSs processed a mixed class of FSs using eye-tracking paradigms. Each FS (*met the deadline by the skin of his teeth*) and a matched non-FS with the same final word (*met the dentist who looked at his teeth*) were embedded in short stories. Both NSs and NNSs fixated on the FS final words fewer times than non-FS final words, while NSs' fixation durations on FS final words were shorter than non-FS final words. The authors claimed that both

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

NSs and NNSs retrieved FSs holistically, but that reading time could only be lessened with the full acquisition of FSs. The authors also pointed out that one possible cause of these conflicting results between fixations and reading times was that NNSs did not process all types of FSs holistically and that a subset of FSs may be processed analytically, thereby increasing overall fixation durations. To avoid the potential issue caused by the heterogeneity of FSs, Conklin and Schmitt (2008) used a similar experimental design but only focused on idioms. In contrast to the previous findings, reading times collected in a self-paced reading task showed that both NSs and NNSs processed idioms significantly faster than control phrases, regardless of whether the context favored a figurative or a literal reading. Siyanova-Chanturia et al. (2011) adopted Conklin and Schmitt's (2008) design, using eye tracking to investigate NSs and NNSs processing of idioms in literal contexts versus figurative contexts. Their findings for NSs replicated the previous study although NNSs processed idioms and novel phrases at a comparable speed. Moreover, figurative readings of idioms were processed more slowly than their literal readings by NNSs. Analysis of fixation time spent before and after the idiom key (the word that determines when an idiom can be recognized as an idiom; Cacciari & Tabossi, 1988) revealed that NNSs spent a longer time on figurative reading before reaching the idiom key, which was the cause of their slowed processing. Similarly, Cieślicka (2006) and Cieślicka and Heredia (2011) found that NSs showed processing advantages for the figurative meanings of idioms, while NNSs showed processing advantages for their literal meanings of idioms. Based on this pattern, Cieślicka (2006) proposed the literal salience hypothesis, that is, the literal meaning of an idiom is more salient to NNSs. However, in a more recent study, Van Ginkel and Dijkstra (2020) used primed lexical decision paradigms to investigate similar questions and found that both NSs and NNSs responded faster to words either figuratively or literally related to primed idioms than unrelated words. Among all these studies, only Siyanova-Chanturia et al. (2011) used NNSs' familiarity ratings to ensure that NNSs knew the target idioms and did not find any advantages for NNSs' processing of FSs.

Studies investigating isolated FSs also generated conflicting findings. Jiang and Nekrasova (2007) conducted grammaticality judgment tasks to investigate the online processing of nonidiomatic FSs (*to tell the truth*) and their matched nonformulaic phrases (*to tell the price*) by English NSs and NNSs. Both NSs and NNSs responded to the FSs faster and with fewer errors than to the nonformulaic controls. Based on these findings, the authors claimed that FSs are holistically represented in both NSs' and NNSs' lexicons. However, focusing on a different type of FS, Gyllstad and Wolter (2016) found the opposite patterns for lexical collocations. Both NSs and NNSs judged collocations more slowly than free combinations, which might be due to the semi-transparent nature of collocations according to the authors. The processing speed was found to be sensitive to the phrasal frequency.

Additionally, addressing the issue of formulaic advantages but with attention given to the fixedness of FSs, different studies have found different patterns for NNSs. Siyanova-Chanturia et al. (2011) used eye tracking to compare the processing of binominal collocations (*bride and groom*) and their reversed forms (*groom and bride*). They found that both NSs and advanced NNSs showed processing advantages for binominals over their reversed controls and attributed the pattern to

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

the high phrasal frequency effect. Vilkaitė (2016) and Vilkaitė and Schmitt (2019) compared NSs and NNSs processing of adjacent collocations (*provided information*) and nonadjacent collocations (*provide some of the information*) also using eye tracking. However, their results showed that unlike NSs, who demonstrated processing advantages for both types of collocations over their novel controls, NNSs only demonstrated advantages for adjacent collocations. In addition, the authors found that NNSs' processing speed was correlated with their pre-existing vocabulary knowledge.

The aforementioned studies mainly concentrated on the comparison between FSs and their nonformulaic controls. Another line of research has focused on internal differences between subcategories of a single type of FS and has mainly concerned the formulaic transfer from L1 to L2 by examining NNS processing of L1-L2 congruent versus incongruent collocations (Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010) or idioms (Carrol & Conklin, 2014, 2017; Carrol et al., 2016). Their major contribution is the finding that congruent FSs have a processing advantage over incongruent FSs, and that this congruency effect may be attributed to the cross-language lexical activation combined with the frequency and compositionality of the FSs. In addition to the consistent finding of the congruency effect, some studies have also found that NNSs' knowledge about FSs might be another influential factor. For example, Wolter and Gyllstad (2011) found that NNSs' online processing patterns paralleled their quality of knowledge patterns. Carrol et al. (2016) found that when L1 knowledge was not available, how fast L2 FSs were processed was significantly related to how familiar they were to the NNS participants. Carrol and Conklin (2017) also reported an emerging familiarity effect in L2 idiom processing. Those studies have demonstrated that NNSs' online processing of FSs was regulated to different extents by how well they knew the FSs. This raises the question of whether the inconsistent processing patterns found for NNSs in previous studies are because some research assumed that NNSs knew the tested FSs but did not test this knowledge.

Indeed, only a handful of studies have examined how well NSs and NNSs know FSs. In a rating study, Carrol et al. (2018) asked English NSs and NNSs to rate idioms' familiarity, transparency, compositionality, and meaning (selecting the correct figurative meaning for an idiom). They found that familiarity had a significant effect on perceptions of transparency and that meaning was strongly affected by compositionality. Based on the NS–NNS differences, the authors concluded that NNSs are more inclined to undertake analytical processing, allowing them to see possible connections between constituent words and whole phrases that NSs tended to overlook. Bardovi-Harlig (2009) used an aural recognition task and an oral production task to investigate the relationship between the recognition and production of FSs by NSs and NNSs, finding that recognition of FSs is a necessary but not sufficient condition for NNSs to correctly produce them. Other factors, such as the degree of familiarity and overuse of some high-frequency expressions, may also cause NNSs to use FSs less frequently than NSs. Nekrasova (2009) conducted a gap-filling task and a dictation task to compare NSs' and NNSs' knowledge of two types of FSs, discourse-organizing bundles (*what do you think*) and referential bundles (*one of the most*). In the gap-filling task, intermediate NNSs' knowledge of FSs was significantly poorer than that of NSs and advanced NNSs. In the dictation task, advanced NNSs

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

outperformed both intermediate NNSs and NSs. In both tasks, discourse-organizing bundles were found to be better acquired than referential bundles by all three groups. The results indicated that speakers' knowledge about FSs was more affected by linguistic registers and discourse functions than by frequency. Based on these findings, the author also suggested that not all types of FSs have the same psycholinguistic status. Carrol and Conklin (2020), who found that the processing of FSs is type sensitive for NS, made the same claim. However, to our knowledge, no study has directly compared NSs' and NNSs' processing of different types of FSs.

The present study

This study set out to compare two subtypes of Chinese FSs, idioms and nonidiom FSs. Idioms (e.g., *kick the bucket*) are fixed phrases whose conventional meanings are not always derivable from the literal meanings of the constituent words. Chinese idioms also include elements of ancient Chinese grammar and lexis (e.g., 一目了然 one-eye-clear-understand *be apprehended at a glance*; a noun "eye" can serve as a verb "look" in ancient Chinese) that can set them apart from nonidiom phrases that conform to modern Chinese grammar, such as lexical bundles (e.g., 一看就懂 yí-kàn-jiù-dǒng one-look-then-understand *be apprehended at a glance*) or collocations (e.g., 重要手段 zhòngyào-shǒuduàn important-method *important method*), even if every constituent word of an idiom identifiably contributes to the overall meaning. Nonidiom FSs (e.g., *to begin with*) in this study refer to fully transparent multiword expressions that "occur as phrases and as coherent semantic units at a relatively high frequency" (Jiang & Nekrasova, 2007, p. 433). On the one hand, the meaning of nonidiom FSs can be derived from the combination of each constituent word's literal meaning, setting them apart from idioms. On the other hand, despite being fully compositional in semantics, nonidiom FSs are different from novel phrases (e.g., *to dance with*) because they enjoy a higher frequency of reoccurrence in texts. Although both idioms and nonidiom FSs are recurrent in language and found to have processing superiority (Wray, 2002), they differ in many linguistic dimensions, such as degree of fixedness and figurative meaning, making idioms intuitively more likely to be processed as holistic units than nonidiom FSs like corpus-derived lexical bundles (Boers & Lindstromberg, 2012). Based on these differences, researchers have adopted the theoretical view that FSs are on a continuum (Coulmas, 1994; N. Ellis, 2012; Wray & Perkins, 2000) that includes "at the one extreme, idiomatic and immutable strings, . . . , and, at the other, transparent and flexible ones containing slots for open class items" (Wray & Perkins, 2000, p. 1). The targets of the present study, idioms and nonidiom FSs, are far apart on the FS continuum, differing mainly in the dimension of idiomaticity. We want to examine whether this difference in the speaker-external (linguistic) dimension also appears in any speaker-internal (processing) dimension for NSs and NNSs.

To achieve this goal, two concurrent data sources, RT and TA verbalizations, are triangulated. RTs, often used to investigate online processing (Jiang, 2011), are an indicator of how much cognitive effort individuals make to process language. TAs, as a window into the minds of speakers, can be used to probe speakers' depth of processing and understanding (Adrada-Rafael, 2017; Bowles, 2010). Leow et al.

(2014) suggested that RT and TA can complement one another in providing a fuller picture of L2 processing. In this study, both are used to answer the following research questions (RQs).

RQ1. Are the two types of FSs processed differently by NSs and NNSs?

RQ2. Do NSs and NNSs display the same quality of understanding (QOU) about the two types of FSs?

RQ3. Are NSs' and NNSs' processing affected by their QOU about the two types of FSs?

Method

Participants

The twenty NSs were Chinese undergraduate and graduate students (12 females; 8 males; $\text{Mean}_{\text{age}} = 27.5$). The NNS participants were 22 Chinese degree learners (12 females; 10 males; $\text{Mean}_{\text{age}} = 22.5$) recruited from four Chinese universities. They came from nine countries: Egypt, Japan, Kazakhstan, Korea, Mongolia, Nepal, Thailand, Russia, and Vietnam. All NNS participants had passed ***Hanyu Shuiping Kaoshi*** (Chinese Proficiency Test) level 6[1] within the last 2 years.

Materials

The FSs used in this study consisted of 48 idioms and 48 nonidiom FSs in Chinese, chosen using the following criteria. First, 76 three-character (3-C) and four-character (4-C) idioms were selected from *The Contemporary Chinese Dictionary* (6th edition) based on their corpus frequency. Second, 33 experienced Chinese second language (CSL) teachers were invited to rate how many HSK-6 learners would know these idioms using a 5-point scale (5 = all HSK-6 learners should know; 1 = no HSK-6 learners should know). Idioms that received an average rating lower than 4 were excluded. Finally, an equal number ($n = 24$) of 3-C idioms and 4-C idioms were selected for the test list. Every idiom was matched with a nonidiom FS (see Table 1 for examples). Because Chinese idioms contain ancient syntactic structures, it was unlikely to find a grammatical nonidiom FS by just changing

Table 1. Examples of test stimuli

FS condition		FS target	Log10 frequency	Stroke *n*
Length	*Type*			
3-C	Idiom	走后门/walk-back-door/*pull strings*	7.18	16
	Nonidiom	走出门/walk-out-door/*walk out of the door*	6.90	15
4-C	Idiom	大吃一惊/big-eat-one-surprise/*be astounded at*	7.51	21
	Nonidiom	大吃一顿/big-eat-one-meal/*eat a big meal*	7.23	20

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

one word in the idiom, as some previous studies have done. However, the nonidiom FSs and their idiomatic counterparts were matched in the following aspects[2]: 1) they had an equal number of characters and similar structures, 2) they shared at least one identical keyword, and 3) they had similar total stroke numbers and whole-phrase frequency. Because some of the nonidiom FSs were absent from the existing Chinese frequency corpora, we followed Libben and Titone's (2008) practice, using the log-transformed page counts of a Chinese website search engine (www.baidu.com) to represent the whole-phrase frequency. Independent samples t-tests ($\alpha = .0125$) showed that frequency and stroke numbers were matched for idiom and nonidiom FSs. Specifically, there were no significant differences in the frequency ($t = -0.57$, $df = 46$, $p = .57$) or stroke number ($t = 0.36$, $df = 46$, $p = .72$) between 3-C idioms and 3-C nonidioms or in the frequency ($t = 0.09$, $df = 46$, $p = .93$) or stroke number ($t = 0.08$, $df = 46$, $p = .94$) between 4-C idioms and 4-C nonidioms.

In summary, the test stimuli consisted of four classes of FSs: 3-C idioms, 4-C idioms, 3-C nonidioms, and 4-C nonidioms. Each class had 24 items (full list in Supplementary Materials). Another 96 ungrammatical phrases were included as filler items. All test items were evenly divided into two counterbalanced blocks (A and B). The block that contained an idiom did not contain its matched nonidiom FS. All NNS participants were given a character list to study at home and then completed a character quiz before the main test began. This procedure was performed to confirm that all characters in the test stimuli were known to NNSs. Two NNSs who did not get 100% correct were removed.

Test instruments

The instruments used to assess the processing and comprehension of FSs are phrase acceptability judgment tasks (AJTs), with one conducted silently (to gather RT) and another conducted with TAs (to collect TA data). Although silent AJTs have been widely used to measure L2 acquisition, researchers have found that NNSs can be inconsistent in making dichotomous judgments (R. Ellis, 1991) and have suggested that the cognitive recourse that L2 learners use for immediate recognition "does not necessarily imply language acquisition" (Leow, 1993, p. 334). Thus, while the silent AJT taps into participants' speeded recognition, TA AJT is a complementary measure that allows us to gather qualitative data about how learners process the FSs. All participants performed the two AJTs in a counterbalanced order with a 1-week interval between the two. In each AJT session, participants saw both blocks of test material with a 5-min break in between. Figure 1 presents the experimental procedure.

Silent acceptability judgment task

In the silent AJT, participants had a brief instruction session followed by a 10-trial training session. First, a fixation cross appeared in the center of the screen for 800 ms and disappeared. A phrase was then exposed in the same position. Participants were asked to judge whether the FS was likely to be used in Chinese by pressing A for "YES" and L for "NO." No time limit was set for each trial, so the FS remained on the screen until a response was entered. However, the test was speeded because

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

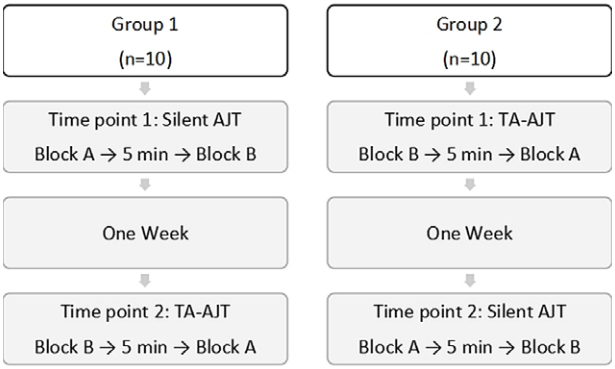

Figure 1. Experimental Procedure.

participants were instructed to make a judgment as quickly as possible. The experiment was run in Paradigm (Perception Research Systems, 2007) on a Lenovo laptop.

Think-aloud–acceptability judgment task

The TA-AJT session was similar to the silent session except that instead of pressing a button to respond participants were instructed to verbalize their Yes or No judgment and report what they were thinking when they made the judgment. Before the experiment, written instructions were given to participants to ensure that they understood they should verbalize whatever thoughts went through their minds when performing the AJT. Participants were also given spoken instructions regarding how to think aloud. First, they were informed that one of the research goals was to obtain a realistic representation of how they understood language. Therefore, they were asked first to read aloud the FS they saw and then judge whether the FS was likely to be read or heard in Chinese. The instructions emphasized that it was important to "speak whatever can help you make a judgment for a stimulus" without worrying about giving explanations, using examples or using incomplete sentences. Both NSs and NNSs were asked to think aloud in the target language, Chinese, since Adrada-Rafael and Filgueras-Gómez (2019) found no qualitative difference between TAs that intermediate NNSs completed in their L1 versus in their L2. As in Kim and Bowles (2019), this was also practical for the researchers given that the NNSs came from nine different L1 backgrounds. Throughout the experiment, an assistant sat beside the participant and prompted them when they fell silent (Bowles, 2010). The whole TA-AJT session was audio-recorded using Audacity and transcribed. A full transcription of an NNS participant is provided (Supplementary Materials).

Coding

The Yes/No judgments about the FSs were first coded for accuracy (correct vs. incorrect) and then for QOU. Previous research (Cooper, 1999) has shown that participants can utilize a variety of strategies in the process of understanding an FS.

Table 2. Operationalization of QOU

	Incorrect	Partially correct	Fully correct
Criteria	Participants show wrong or no knowledge of the FS	Participants show partial knowledge of the FS	Participants show full knowledge of the FS
Evidence	○ Acknowledge the FS has been "heard/seen/learned" or identify the FS to be a certain type of expression but admit "have forgotten/have no idea what it means" ○ Judge a correct FS to be an incorrect one and provide a "correction" ○ Provide an interpretation/example that shows wrong knowledge of the FS ○ Provide a wrong metalinguistic analysis	○ Provide a literal interpretation/example for an idiom whose figurative meaning is the default use[3] ○ Provide a metalinguistic analysis that does not show full understanding ○ Provide a related but inaccurate interpretation/example ○ Provide an example with a correct context but not exactly correct grammar/pragmatics	○ Provide a "cliché" type answer, stating: "it's just what I often say", or use an "A just means A" sentence to imply there is nothing worth saying ○ Provide an example with correct context, grammar, and pragmatics ○ Provide a correct interpretation (a figurative interpretation for an idiom) ○ Provide a metalinguistic analysis that can show full understanding

Because the TA procedure in this study only intended to probe participants' success in understanding FSs rather than the strategies they used, we coded three levels of QOU based on previous research (Boers & Demecheleer, 2001; Schmitt et al., 2001): incorrect, partially correct, and fully correct. Table 2 presents the operationalization of the coding procedure, adapted from previous coding procedures for depth of processing (Adrada-Rafael, 2017; Leow & Mercer, 2015).

The first author and a research assistant coded 25% of the data independently. The interrater agreement was 100% for judgment accuracy and 94.6% for QOU. After discussing some controversial cases, the two raters coded another 5% of the data for QOU, and the interrater agreement reached 98.4% (Cohen's kappa = 0.85). The results were considered high enough for the first author to code the remaining data alone.

Results

Test order effect

To examine whether there was a test order effect (silent-TA vs. TA-silent), judgments and RTs from the silent AJT sessions (time point 1 for Group 1 and time point 2 for Group 2) were compared using independent samples *t*-tests ($\alpha = .0125$). No significant differences in judgment accuracy were found between

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

the two orders for either NSs ($p = .75$) or NNSs ($p = .29$). There were no significant differences in mean RTs based on test order for either NSs ($p = .49$) or NNSs ($p = .21$). Therefore, the two order conditions were merged, with (non)-nativeness being the only between-group condition in the following analyses.

Preliminary analysis

The two AJTs resulted in 1920 datasets (20 participants × 96 target items) from each group. Each dataset includes four data points: a silent Yes/No judgment and its RT and a verbalized Yes/No judgment and its TA report. Data were trimmed by removing whole datasets instead of single data points. First, only consistently correctly judged items in the two AJT sessions were included in the analyses. This procedure removed 2% of the NS datasets and 17% of the NNS datasets (Table 3). To further trim outliers, a participant's RT data that were 3 standard deviations from their mean were eliminated. This procedure removed another 2.7% of the NS datasets and 1.6% of the NNS datasets. After removing these data, 1830 sets of NS data (95%) and 1564 sets of NNS data (81%) were retained for analysis. RT data were then Log10-transformed to reduce skewing. Log10RT data were analyzed using linear mixed-effects models (LMMs), and QOU coding was analyzed using generalized linear mixed-effects models (GLMMs) in R (version 4.1.0; R Development Core Team, 2021) using the lmerTest package (Kuznetsova et al., 2017). When fitting the models, Group, Type, and Length (or a subset of these three factors) were included as fixed effects, and subjects and items were entered as random effects (Baayen et al., 2008). The maximal random effects structure was first included and then reduced when the model failed to converge, following Bates et al.'s (2015) advice. Pairwise comparisons were computed using the emmeans package (Lenth et al., 2021). In all analyses, we report the model structure and the significance of the fixed effects, including the coefficient (β), standard error (SE), and *t* value (*z* value for QOU). The full model outputs are provided in the Supplementary Materials (Tables C1-E2).

Table 3. Judgment error rate and consistently correct rate in two AJTs

		NS			NNS		
		Error rate		Consistently correct rate	Error rate		Consistently correct rate
FS condition		Silent	TA		Silent	TA	
Length	*Type*						
3-C	Idiom	1.04%	1.87%	97.92%	16.67%	15.21%	78.96%
	Nonidiom	1.25%	1.04%	98.13%	11.04%	5.21%	87.08%
4-C	Idiom	0.00%	0.00%	100.00%	11.67%	12.08%	81.67%
	Nonidiom	2.50%	2.50%	95.83%	13.13%	8.54%	84.38%

RQ1. Online processing

RQ1 asked whether NSs' and NNSs' online processing patterns were different. Raw RT data were first analyzed descriptively (Table 4). For NSs, the mean RTs of both 3-C and 4-C idioms were shorter than those of their nonidiom counterparts. For NNSs, the mean RTs of both 3-C idioms and nonidioms were shorter than their 4-C counterparts; the mean difference between 3-C and 4-C FSs was greater for idioms (357 ms) than nonidioms (114 ms). The mean RT of 4-C idioms was the shortest in the NS group but the longest in the NNS group.

The LMM analysis was then conducted for Log10RTs with Group, Type, and Length entered as fixed effects and subjects and items entered as random effects including random intercepts for subjects and items and by-subject and by-item random slopes for Group, Type, and Length. The analysis returned a significant effect for Group ($\beta = 0.23$, $SE = 0.04$, $t = 6.36$, $p < .00$), Type ($\beta = 0.03$, $SE = 0.01$, $t = 2.20$, $p = .03$), and the Group × Length interaction ($\beta = 0.08$, $SE = 0.02$, $t = 4.04$, $p < .00$) but not for Length ($\beta = -0.01$, $SE = 0.01$, $t = -0.83$, $p = .41$), the Group × Type interaction ($\beta = -0.01$, $SE = 0.02$, $t = -0.32$, $p = .75$), the Type × Length interaction ($\beta = 0.00$, $SE = 0.02$, $t = -0.10$, $p = .92$), or the three-way interaction ($\beta = -0.04$, $SE = 0.02$, $t = -1.62$, $p = .11$). To determine whether a significant Type and Length effect existed for each group, the LMM analysis was conducted separately for NSs and NNSs. For NSs, the Type effect was significant ($\beta = 0.03$, $SE = 0.01$, $t = 2.60$, $p = .01$), with both 3-C and 4-C idioms processed faster than their nonidiom counterparts ($ps < .05$), but the Length effect was nonsignificant ($\beta = -0.01$, $SE = 0.01$, $t = -0.82$, $p = .42$). In contrast, for NNSs, the Type effect was nonsignificant ($\beta = 0.03$, $SE = 0.02$, $t = 1.27$, $p = .21$), but the Length effect was significant ($\beta = 0.07$, $SE = 0.01$, $t = 4.75$, $p < .001$), with 3-C idioms and nonidioms both processed faster than their 4-C counterparts ($ps < .05$). The overall pattern demonstrated that NSs' processing was more likely to be affected by the type of FS, whereas NNSs' processing was more likely to be affected by the length of FSs.

Table 4. Descriptive statistics of RTs (ms)

FS condition		NS			NNS		
		Mean	*SD*	*n*	Mean	*SD*	*n*
Length	*Type*						
3-C	Idiom	1041	585	458	1841	1173	371
	Nonidiom	1119	558	459	1961	1246	409
4-C	Idiom	987	419	474	2198	1348	385
	Nonidiom	1085	521	439	2075	1220	399

RQ2. Quality of understanding

RQ2 asked whether there was a difference in NSs' and NNSs' QOU with different types/lengths of FSs. To answer this question, the QOU coding was analyzed. As

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

Table 5. Frequency of QOU

FS condition		NS			NNS		
		Incorrect	Partially correct	Fully correct	Incorrect	Partially correct	Fully correct
Length	*Type*						
3-C	Idiom	0.00%	2.84%	97.16%	3.24%	19.46%	77.30%
	Nonidiom	0.22%	0.66%	99.13%	0.98%	13.20%	85.82%
4-C	Idiom	0.21%	0.42%	99.37%	10.62%	19.17%	70.21%
	Nonidiom	0.68%	0.45%	98.86%	3.26%	20.80%	75.94%

Table 5 illustrates, NSs' incorrect TAs were observed less than 1% of the time under different FS conditions, and partially correct TAs were less than 3%. However, the frequencies of NNSs' incorrect TAs varied greatly under different FS conditions, ranging from 1.0% (3-C nonidioms) to 10.6% (4-C idioms), as did the frequencies of partially correct TAs, ranging from 13.2% (3-C nonidioms) to 20.8% (4-C nonidioms). The most successful FSs in the NS group were 4-C idioms with the highest fully correct ratio (99.4%), which were the most unsuccessful FSs in the NNS group with the lowest fully correct ratio (70.2%).

The GLMM analysis was first conducted on QOU with Group, Type, and Length being entered as fixed effects and subjects and items added as random effects including a maximal random effects structure. After a series of model fitting following a backwards elimination algorithm, the model with minimal random effects structure still failed to reach convergence[4]. The fixed-effect-only generalized linear model finally converged and returned a significant effect only for Group ($\beta = -0.12$, $SE = 0.05$, $z = -2.42$, $p = .02$). All the other effects including Type ($\beta = 0.01$, $SE = 0.05$, $z = 0.19$, $p = .85$), Length ($\beta = -0.01$, $SE = 0.05$, $z = 0.22$, $p = .83$), and the interaction of Group × Type ($\beta = 0.05$, $SE = 0.07$, $z = 0.72$, $p = .47$), Group × Length ($\beta = -0.01$, $SE = 0.07$, $z = -1.33$, $p = .19$), Type × Length ($\beta = -0.01$, $SE = 0.07$, $z = -0.21$, $p = .84$), and Group × Type × Length ($\beta = 0.03$, $SE = 0.10$, $z = 0.32$, $p = .75$) were all nonsignificant. To explore whether Type and Length had an effect on QOU within each group, separate analyses were conducted with the same model fitting procedure. The analysis for NSs yielded no significant effect for Type ($\beta = 0.00$, $SE = 0.05$, $z = 0.19$, $p = .85$), Length ($\beta = 0.01$, $SE = 0.05$, $z = 0.22$, $p = .83$), or their interaction ($\beta = -0.01$, $SE = 0.07$, $z = -0.21$, $p = .84$); pairwise comparisons also returned no statistical significance (ps > .05). For NNSs, although the main analysis did not show a clear effect for Type ($\beta = 0.06$, $SE = 0.05$, $z = 1.12$, $p = .26$), Length ($\beta = -0.09$, $SE = 0.06$, $z = -1.54$, $p = .12$), or their interaction ($\beta = 0.02$, $SE = 0.08$, $z = 0.24$, $p = .81$), pairwise comparisons showed that the difference between two lengths was significant ($p = .04$) with 3-C FSs understood better than 4-C FSs, and the difference between two types was marginally significant ($p = .07$) with nonidioms better understood than idioms. The overall results show that NSs' understanding of FSs with different types and lengths did not vary much. However, NNSs' understanding of different types

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

and lengths of FSs varied to different extents. Longer FSs were understood more poorly than shorter FSs; the QOU of idioms was lower than that of nonidioms. Although inferential statistics showed only a marginal difference in NNSs' understanding of idioms versus nonidioms, the frequency distribution demonstrated that both 3-C and 4-C idioms elicited more incorrect TAs and fewer fully correct TAs than their nonidiom counterparts. Additionally, NNSs' relatively high frequencies of incorrect TAs (4.5%) and partially correct TAs (18.2%) indicated that although NNSs' judgments of the FSs were repeatedly correct, their FS knowledge might be partial or completely wrong.

RQ3. Effect of the QOU on online processing

To determine whether the QOU has an effect on the online processing of FSs, the LMM used in RQ1 was conducted again on Log10RT with QOU added as the covariate. The results (Table 6) generally replicated the findings of RQ1. However, no statistical effect of QOU was obtained. To examine if there was any interaction between QOU and Type and Length variables, separate LMM analyses were run for NSs and NNSs. Type and Length were combined into Category (4 levels: 3-C idiom, 4-C idiom, 3-C nonidiom, and 4-C nonidiom) to remedy the multicollinearity issue. Results (Table 7) show that QOU was not a significant predictor of either NSs or NNSs online processing. Pairwise comparisons (Table 8) only yielded a marginal difference in NNSs' processing speed of fully correct versus partially correct 3-C nonidioms, but no reliable effect of QOU on the processing speed was found for any other Category condition in either group. This pattern of results suggests that NSs and NNSs' online processing of a specific category of FSs was not directly affected by how well they understood the FSs.

Table 6. Results of fixed effects of the LMM analysis with QOU as the covariate

	β	*SE*	*t*
Fixed effects			
(Intercept)	3.01	0.02	135.68***
Group: NNS	0.23	0.04	6.19***
Type: Nonidiom	0.03	0.01	2.33*
Length: 4-C	−0.01	0.01	−1.16
QOU: Partially correct	−0.02	0.02	−1.32
QOU: Fully correct	−0.05	0.01	−3.43***
Group (NNS) × Type (Nonidiom)	0.00	0.02	0.07^{+}
Group (NNS) × Length (4-C)	0.08	0.02	4.64***
Type (Nonidiom) × Length (4-C)	0.00	0.02	−0.10
Group (NNS) × Type (Nonidiom) × Length (4-C)	−0.04	0.02	−1.58

*$p < .05$; ***$p < .001$; $^{+}p = .05–.10$.

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

Table 7. Results of fixed effects of LMM analyses for NS and NNS group

	NS group			NNS group		
	β	*SE*	*t*	β	*SE*	*t*
Fixed effects						
(Intercept)	2.91	0.09	31.24***	3.09	0.04	85.39***
Category2	0.06	0.18	0.34	0.34	0.03	12.23***
Category3	0.04	0.18	0.20	−0.04	0.05	−0.80
Category4	0.03	0.01	2.45*	0.36	0.03	10.49***
QOU1	0.11	0.10	1.07	−0.02	0.03	−0.68
QOU2	0.07	0.09	0.75	−0.03	0.03	−1.29
Category2 × QOU1	0.03	0.22	0.12	0.02	0.03	0.63
Category3 × QOU1	0.00	0.21	0.00	0.06	0.05	1.11
Category4 × QOU1	−0.12	0.12	−0.98	−0.01	0.04	−0.34
Category2 × QOU2	−0.07	0.18	−0.38	0.03	0.03	0.94
Category3 × QOU2	0.00	0.18	−0.03	0.04	0.05	0.85
Category4 × QOU2	n/a[a]	n/a	n/a	0.00	0.03	−0.07

Note. These labels are used in Tables 7 and 8.
Category2: 4-C idiom; Category3: 3-C nonidiom; Category4: 4-C nonidiom; QOU1: partially correct; QOU2: fully correct.
[a]n/a was due to the zero instances of incorrect QOU for 3-C idioms in the NS group.
**p* < .05; ****p* < .001.

Table 8. Pairwise comparisons for QOU × Category

	t value (NS group)				*t* value (NNS group)			
	3-Idiom	4-Idiom	3-Nonidiom	4-Nonidiom	3-Idiom	4-Idiom	3-Nonidiom	4-Nonidiom
Contrasts								
QOU0–QOU1	n/a	−0.71	−0.60	0.06	0.68	−0.11	−0.89	1.20
QOU0–QOU2	n/a	−0.00	−0.41	−0.75	1.29	0.38	−0.22	1.43
QOU1–QOU2	0.89	1.22	0.48	−0.70	1.29	0.64	2.36$^+$	0.40

^+p = .05–.10.

Discussion

Research question 1

The first question asked whether NSs and NNSs' FS processing patterns were different. The RT data show that NSs' processing was more likely to be modulated by the type of FS, with idioms being processed faster than nonidiom FSs regardless of

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

length. In contrast, NNSs' processing was more likely to be modulated by length of FS, with 3-character FSs being processed faster than 4-character FSs regardless of type.

The finding that NSs and NNSs demonstrated different processing patterns when reading idioms versus literal phrases replicates some past results[5] (Carrol & Conklin, 2014, 2017; Siyanova-Chanturia et al., 2011). The NS processing pattern may indicate that idioms and nonidiom FSs have different psychological statuses in NSs' mental lexicons. In this study, phrasal frequencies were matched for idioms and nonidiom FSs; therefore, this pattern could also suggest that idiomaticity rather than phrasal frequency plays a more significant role in modulating NSs online processing of FSs, a phenomenon known as the *idiom superiority effect* (Tabossi et al., 2009). Regarding how idiomaticity could modulate processing speed, we speculate two possibilities. The first reason might involve compositionality. Idioms are less compositional than nonidioms, and some parts of an idiom's constituents may not identifiably contribute to its meaning. For Chinese idioms, lower compositionality could also mean that contemporary Chinese speakers have trouble parsing an idiom's ancient syntactic configuration even if all the constituent words are to some extent related to the whole idiom's meaning. Thus, the acquisition of idioms may involve an item-based mechanism, requiring storing each form-meaning association in the lexicon. Therefore, once a target FS was recognized as an idiom, its meaning would be directly retrieved just as with long complex words (Gibbs & O'Brien, 1990; Swinney & Cutler, 1979), hence the processing advantage. The second reason concerns how fast an idiom can be recognized as an idiom, which might have to do with its formal fixedness (Fraser, 1970; Swinney & Cutler, 1979). This was evidenced by the internal processing difference between 4-character and 3-character idioms, with 4-character idioms processed faster than 3-character idioms (but not significantly so). Chinese 4-character idioms, also referred to as chéng-yǔ (proverbs; literally translated as "fixed language"), are mostly obtained from classical Chinese and have absolutely frozen forms that are not subject to any lexical substitution or syntactic operation. Because of their fixed nature, we suspect that NSs were able to recognize a 4-character idiom as an idiom at a superior speed. However, we could not rule out the possibility that the 4-character idioms were also less compositional than 3-character idioms and thus more likely to be prestored and retrieved directly. To distinguish between these possibilities, a post hoc analysis was conducted using NSs' ratings on Chinese idioms' decomposability in a large-scale norming study (Zheng, 2019). The analysis showed that the target 4-character idioms were rated more decomposable than the 3-character idioms ($t{=}-4.58$, $p < .00$, ratings listed in Test Stimuli in Supplementary Materials). This result suggests that the faster processing speed of 4-character idioms was more likely due to the form fixedness. Even though 4-character idioms were longer and more decomposable than 3-character ones, NSs can quickly predict the whole sequence without having to finish reading every word. The absence of the length effect in the NS group could be explained by the perceptual span of Chinese NSs. Since 3- and 4-character phrases fall in the average perceptual span of native Chinese readers (Inhoff & Liu, 1997), the one-character difference may have been too subtle to be reflected in the RT data, especially given the ceiling effect.

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

No idiomaticity advantage was observed for NNSs. This result indicates that idiom and nonidiom FSs appear to be processed in a similar fashion by NNSs. Furthermore, the length effect might suggest that it is very likely that NNSs processed both idiom and nonidiom FSs in a word-by-word manner. This verbatim processing of idioms by NNSs suggests that NNSs and NSs handle idioms differently. Such a difference can more clearly be seen through the processing of 4-character idioms, the fastest type for NSs to grasp but the slowest type for NNSs to grasp. These results reinforce the claim (Abel, 2003; Myles & Cordier, 2017) that just because an FS is stored holistically in NSs' lexicons does not mean that it is also stored holistically in NNSs' lexicons. For nonidiom FSs, the presence of a length effect showed that they may also be processed in a verbatim fashion by NNSs. In contrast, in Jiang and Nekrasova's (2007) study where nonidiom FSs (mostly lexical bundles, e.g., *to start with*) were found to be processed faster than nonformulaic controls (e.g., *to bring with*) by NNSs, the authors claimed that NNSs processed nonidiom FSs holistically as single units. Because the current study did not have nonformulaic controls, the two studies may not be directly comparable. Nevertheless, the unconformity prompts us to reconsider whether the speed advantage can be directly taken as an account for the "holistic" nature, the question previously raised by Siyanova-Chanturia (2015) and Siyanova-Chanturia and Martinez (2015). This study showed that this was not the case, at least for NNSs. As cautioned by Siyanova-Chanturia (2015), the speed of whole phrase processing can tell little about whether or not individual constituents are processed and thus cannot be used as evidence to ascertain analytical or holistic processing.

It is also worth mentioning that although the length effect implied that NNSs may approach both types of FSs in a word-by-word manner, the silent AJT was simply a visual recognition task. What can be inferred from the RT data may be more pertinent to how an FS's form was recognized than how its meaning was processed.

Research question 2

The second question inquired whether there is any difference in speakers' QOU about different types of FSs and was addressed using TA data. The results show that NSs were at or near the ceiling for every subtype of FS. This is highly predictable given that the selected FSs are all frequently used in written or spoken Chinese and hence very familiar to adult NSs. The high familiarity was also evidenced by the fact that approximately half of NSs' TAs (54.5% of idioms, 49% of nonidioms) were "cliché" type responses, such as "It's just what I often say." Clearly, NSs tend to use their intuition to make judgments, which is often considered a shallow form of processing that involves little cognitive effort (Leow & Mercer, 2015). The small proportion of incorrect TAs (0.3%) from NSs occurred not because of incorrect understanding but because they questioned whether a lexical bundle (e.g., 不是所有 not-copula-all-all *not all of*) could be used in isolation.

For NNSs, their understanding of both types of FSs was to some extent limited, especially for the idioms. The different QOU between idioms and nonidioms might be explained by the different contexts in which they commonly occur. Nonidiom FSs are commonly used in daily communications. Thus, NNSs who study

Chinese in China not only have adequate exposure to nonidiom FSs but also abundant opportunities to use them. The usage-based model predicts that maximal and interactive exposure can enhance learners' sensitivity to language (Abbot-Smith & Tomasello, 2006; Bannard & Lieven, 2009). Therefore, nonidiom FSs were better acquired by NNSs. One may argue that since the phrasal frequency of idioms and nonidioms was matched in this study, learners should have an equal chance of being exposed to idioms as they are to nonidioms. As pointed out by Carrol and Conklin (2020), corpus frequency is different from subjective familiarity although the two are often highly correlated. The fact is that idioms are highly frequent in formal written texts, such as newspaper articles and literary works, a register and genre that NNSs seldom encounter, unlike NSs. Thus, despite the matched corpus frequency, NNSs' subjective familiarity with idioms may be lower than with nonidioms.

Another explanation for the better understanding of nonidiom FSs may have to do with compositionality. Nonidiom FSs are fully transparent items whose meaning is the combination of each individual word. Thus, learners can make sense of an unfamiliar nonidiom FS simply by analyzing its vocabulary items (Libben, 1998; Sandra, 1990), as exemplified in the excerpts below.

Excerpt 1

Participant 6 (Korean L1)

Target (4-C nonidiom): 轻松自在/light-slack-self-exist/*relaxed and unrestrained*

TA: "我没见过这个, 但是能明白, 意思就是很轻松, 很自由."

"*I have never seen this one but can understand; it just means very relaxed, very free.*"

The target FS in Excerpt 1 is a nonidiomatic collocation composed of two disyllabic adjectives 轻松 "relaxed" and 自在 "unrestrained." Although the participant claimed that she never used the phrase before, she successfully put together a correct meaning ostensibly by interpreting the two disyllabic words that she knew. However, when learners applied the same strategy to idioms, comprehension errors occurred. Excerpt 2 and 3 demonstrated this process.

Excerpt 2

Participant 9 (Japanese L1)

Target (3-C idiom): 出人命/happen-human-life/death causing

TA: "对的, 就是出生了一个人."

"*Correct, meaning 'to give birth to a person'.*"

Excerpt 3

Participant 11 (Thai L1)

Target (4-C idiom): 谈天说地/talk-sky-speak-earth/*talk of everything under the sun*

TA: "对的, 就是谈论天气."

"*Correct, meaning 'to talk about the weather'.*"

In Excerpts 2 and 3, participants were not familiar with the target idioms, but they attempted to make sense of them by analyzing the literal meanings of the constituent words. Nevertheless, the figurative meaning of an idiom is not always derivable from constituent words (Titone & Connine, 1994), so the strategy led to an incorrect understanding. The processing strategies exhibited in the two excerpts were consistent with what the literal salience hypothesis proposed, that "understanding L2 idioms entails an obligatory computation of the literal meanings of idiom constituent words" (Cieślicka, 2006, p. 115). Therefore, NNSs' lower QOU about idioms might also be caused by the decomposing strategy they used to process idioms.

Regarding NNSs' poorer understanding of longer FSs, NNSs' TAs showed that they often understood the basic meanings of the longer FSs but provided example sentences with grammar or pragmatic errors. This is why 4-character FSs had more partially correct TAs than 3-character FSs. These instances suggest that longer FSs might be more difficult to fully acquire not because they are semantically more complex but because they are contextually more constrained. What the TA data further tell us is that there are some FSs that "learners think they know but they do not" (Laufer, 1989, p. 11). With judgment data alone, one might be tempted to conclude that a learner knows the correct meaning of any FS that s/he consistently judged correctly; however, TA data show when this is and is not the case.

Research question 3

The third question investigated the role that QOU plays in the online processing of FSs. The results suggest that both NSs and NNSs processing speed was not directly related to their understanding of FSs. The finding about NSs is not surprising given the ceiling effect for NSs' FS knowledge (RQ2). The disassociation between NNSs' understanding and processing speed seems counterintuitive, as previous research has shown that the better understood FSs were also processed faster (Liontas, 2003; Vilkaitė & Schmitt, 2019; Wolter & Gyllstad, 2011). However, previous research either did not directly test the understanding of FSs or did not strictly measure online processing. The dissociated relationship may indicate that NNSs employed different sources of knowledge to complete the two tasks. Although processing speed elicited in the silent AJT probes the immediate recognition of the forms of FSs, the QOU elicited in the TA-AJT probes the understanding of the meanings of FSs. Our findings suggest that NNSs need not fully acquire an FS to perform speedy and successful recognition. NNSs' reports show that they can recognize an FS using surface-level knowledge, such as just having "seen/heard it before." It was also observed that NNSs utilized a variety of strategies to compensate for limited knowledge and make a correct judgment. Consider the following three "partially correct" excerpts.

Excerpt 4

Participant 3 (Japanese L1)

Target (4-C idiom): 谈天说地/talk-sky-speak-earth/*talk of everything under the sun*

TA: "对的, 不确定是什么意思, 但汉语里有 '什么天什么地'这样的结构."

"*Correct, not sure about its meaning, but Chinese has the structure like something Sky something Earth.*"

Excerpt 4 shows that the learner did not know the idiom but was aware that Chinese idioms can have a prototypical structure "something A, something B." This instance illustrates that advanced learners who have developed good idiom awareness were able to use metalinguistic knowledge to make a correct acceptability judgment, even though they did not know the idiom's figurative meaning.

Excerpt 5

Participant 16 (Mongolian L1)

Target (3-C idiom): 开夜车/drive-night-car/*burn midnight oil*

TA: "对的. 蒙语也有这个说法. 就是晚上开车."

"*Correct. Mongolian has this saying, too, just meaning 'to drive at night'.*"

Excerpt 5 presents a case of L1-to-L2 transfer, which has been found to facilitate L2 FS processing[6] (Carrol et al., 2016; Irujo, 1986; Yamashita & Jiang, 2010). However, this excerpt shows us that L1 transfer can also be detrimental to L2 idiom comprehension. In this case, idioms in the two languages share the same literal meaning but not the same metaphorical extension.

Excerpt 6

Participant 9 (Korean L1)

Target (4-C idiom): 一帆风顺/one-sail-wind-smooth/*everything is going on smoothly*

TA: "'一路顺风'的话听过, '一帆风顺'应该差不多的意思."

"*I only heard 'one-road-smooth-wind' (another idiom, meaning wish you a happy voyage); 'one-sail-wind-smooth' should have a similar meaning.*"

Excerpt 6 demonstrates the role that pre-existing knowledge plays in idiom processing. The learner judged an unseen idiom based on a known idiom that has a similar form to the target. This shows that advanced learners have stored some L2 idioms in their repertoires and that this knowledge network can provide some clues for NNSs to make a well-formed conjecture about an idiom's figurative meaning (Liontas, 2003).

The above excerpts demonstrate that advanced NNSs were able to employ strategies to make a correct judgment, although these strategies were not sufficient to enable learners to achieve fully accurate comprehension. Conversely, as Underwood et al. (2004) suggested, although it is only with full mastery of FSs that the processing time can be shortened, partial mastery is still rewarded in some way. We speculate that the processing strategies used by NNSs had an offsetting effect, causing partially understood idioms to be recognized as quickly as well-understood nonidioms. However, we caution that this finding was based on a limited number of observations. The issue of how specific L2 processing strategies may contribute to processing speed (see van Gelderen et al., 2004) needs to be accounted for by larger-scale studies.

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

Conclusions, implications, and future research

Through analyzing RT data and TA protocols, this study investigated NSs and NNSs' processing of idiom and nonidiom FSs, comparing whether the two were processed in the same way. Given the relatively small sample size, we consider that the findings constitute tentative rather than conclusive answers to our research questions. For NSs, the RT data indicate that idioms were processed significantly more rapidly than matched nonidiom FSs. This result may serve as evidence that the processing advantage found for individual types of FSs may not imply that all FSs are created equal. Moreover, it provides psycholinguistic evidence for the continuum view of FSs. Because both idiom and nonidiom FSs enjoy some degree of formulaicity, the processing difference found between the two cannot be categorically interpreted as one being prestored and the other not. Rather, it would be better to consider the difference between the two as a gradient, which may have something to do with the extent to which an overall meaning needs to be prestored due to the noncompositionality and the extent of the formal fixedness. For NNSs, what the RT data tell us is that NNSs might very well be utilizing a words-and-rules approach to unraveling both idioms and nonidioms because longer FSs were processed significantly more slowly than shorter ones regardless of type. Underwood et al. (2004) suggested that the processing difference between NSs and NNSs is a product of lifetime exposure to FSs. Boers and Lindstromberg (2012) claimed that even when NNSs encounter FSs a sufficient number of times, they are also less likely to process FSs in a way that resembles NSs. The findings of this study provide support for these claims because the processing patterns of NNSs and NSs were fundamentally different. However, it is worth noting that the RT difference between NSs and NNSs might not be directly used as evidence of holistic versus analytical processing because the whole form recognition task did not tap into how the internal words are processed.

TA data showed that NSs' understanding of FSs reached ceiling while NNSs' understanding of idioms was significantly poorer than that of nonidioms despite the comparable processing speed. What the TA data further reveal is that NNSs demonstrated some explicit idiom awareness, which may have facilitated their recognition of the idioms that they did not fully understand. Kim (2016) argued that the awareness of idioms is a stepping stone to the more effective learning of idioms. We argue that awareness is also a step in the direction of automatic idiom processing. From a methodological perspective, our findings demonstrate that learners' cognitive processes, as revealed by TA data, can be a useful source to investigate learners' awareness (Leow, 2000).

Another methodological issue raised by the comparison of RT and TA data that may merit reconsideration is whether the psycholinguistic account of NNSs' FS processing should be established on the finding of speedy and successful recognition of the target FSs or established on the finding of speedy and successful understanding of the target FSs. Note that NSs have solid knowledge of FSs, and under this premise, the difference in RTs was translated into a different status in the mental lexicon. However, the premise did not hold for NNSs even at advanced proficiency levels. Thus, it might be premature to make claims about how FSs are represented by NNSs without knowing whether they actually knew the FSs or not. As Boers and

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

Lindstromberg (2012) claimed, "one can argue that it is only when a sequence is deeply entrenched in a language user's long-term memory that it qualifies as truly formulaic for that user" (p. 85).

Although the present research did find new evidence that the two subtypes of FSs are qualitatively different, there are some limitations in the research that must be mentioned. First, the chosen idioms are heterogeneous in many ways, such as degree of literality (whether an idiom's literal interpretation can be used in any real-world context). Because the focus of the study is to see whether NNSs and NSs process the same sets of FSs in a similar way, we did not investigate the impact of these linguistic variables. However, these variables may play a different role for NSs than for NNSs (Hubers et al., 2020). Future research efforts on how these variables modulate the ease and difficulty of FS processing can certainly provide insights into the differences observed between NSs and NNs. The same issue also existed in nonidiom materials. The nonidiom FSs used in the study included lexical bundles and collocations. Lexical bundles are "frequently recurring strings of words that often span traditional syntactic boundaries" (Tremblay et al., 2011, p. 569), and collocations that are fully decomposable co-occurrences whose meanings are transparent and unlikely to "cause trouble for L2 learners from a comprehension perspective" (Wolter & Yamashita, 2018, p. 396). Previous research (Carrol & Conklin, 2020) found that different subtypes of nonidiom FSs are processed differently by NSs. Our TA data also revealed that NSs might have more trouble judging lexical bundles than collocations because lexical bundles sound like incomplete sentence fragments. Therefore, to fully account for why different subtypes of FSs are processed differently, the subdivision of nonidiomatic formulae is also needed. Another issue regarding the test material is that the current study did not have a baseline condition, and thus, strictly speaking, any processing advantage found in this study may not be claimed to be the formulaic advantage over the nonformulaic phrases. In addition, although the test materials were selected based on the familiarity ratings of experienced CSL teachers, the selection cutoff (4 out of 5) may have been too lenient because not all the selected idioms were equally familiar to NNS participants. Future research may consider adopting a stricter cutoff to make more conclusive claims about whether FSs have mental representations in the L2 lexicon. Alternatively, Hubers et al. (2020) suggested that the selection of idiom material could be based on L2 speakers' intuition ratings, which have been proven to be reliable and better reflect L2 knowledge than L1 speakers' intuitions.

Although the TA data in this study revealed novel findings that were not reflected in the RT data, the controversial issue of this method, namely its reactivity, is still worth mentioning. In the preliminary analysis, we found that the NNSs' error rate in the TA-AJT session (10.3%) was lower than that in the silent AJT session (13.1%). Closer scrutiny showed that the difference was only due to nonidiom FSs. That is, nonidiom FSs were judged more accurately in the TA session than in the silent session, while the error rates of idioms were similar in the two sessions. Based on this pattern, we speculate that the difference in error rates may be attributable to the read-aloud effect. In the TA session, participants read aloud the stimulus that they saw. This performance helped NNSs avoid some reading errors that could happen in the silent session. Because idioms were not judged more accurately by NNSs in the TA session and neither type of FS was judged more accurately by NSs in the TA session, we concluded that TA protocols did not significantly influence speakers' cognitive processes of reading FSs, which is generally consistent with previous findings (Bowles, 2008, 2010; Bowles & Leow, 2005; Leow & Morgan-Short, 2004). Furthermore, the study demonstrates that TA protocols and RT data can complement each other and provide a fuller picture of L2 processing (Leow et al., 2014). We believe that this methodology may be useful for future research efforts. However, more comprehensive and reliable findings will require larger-scale studies.

In conclusion, the present study compares the processing of two different types of FSs to gain insights into FS processing by NSs and NNSs. The findings demonstrate the importance of tapping into NNSs' thought processes to provide evidence rather than just speculating as to why NNSs' processing or behavioral patterns are different from those of NSs.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0142716421000552

Acknowledgments. We would like to thank Professor Chilin Shih and Misumi Sadler for their feedback on an earlier version of this article as well as Dr. Gareth Carrol and the anonymous AP reviewers for their insightful comments. This research is supported by a Chinese Lexicography Research Center Grant (CSZX-PY-202009).

Notes.

1. The HSK is a standardized Chinese proficiency test that assesses nonnative Chinese speakers' ability to use Chinese in daily life and academic and professional settings. HSK6 is the highest level intended for advanced learners and requires a minimum vocabulary of 5,000 words.

2. Because the focus of this study was to compare idioms and nonidiomatic FSs, the subtypes of nonidioms were not manipulated.

3. In the preliminary test, native speakers tended to interpret idioms' figurative readings, even when their literal readings were highly plausible. Based on this observation, we coded the literal reading of an idiom as partially correct.

4. According to Park et al. (2020), mixed-effect models with two- and three-predictor conditions are more likely to encounter non-convergence issues. We suspected that the non-convergence problem occurring in the QOU analysis might also have to do with the unbalanced distribution of the three QOU levels with the frequency of "fully correct" TAs being extremely high and the "incorrect" TAs extremely low, given that the analyzed data were repeatedly correct items.

5. As the anonymous reviewers pointed out, unlike previous studies, the current study did not have non-formulaic baseline controls. Given that the nonidiom FSs used in this study are also literal in nature, the comparison here only intended to emphasize the different processing patterns found for NSs and NNSs.

6. An anonymous reviewer pointed out that the similarity between a particular L1 and Chinese could lead to an easier interpretation of the FSs. According to previous research, the congruency effect seems to be independent of the similarity between L1 and L2, for example, Japanese and English (Yamashita & Jiang, 2010). Because L1 transfer instances were scarce in this study, we believe that the similarities between participants' first languages and Chinese should not be a confounding factor.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, **23**(3), 275–290. https://doi.org/10.1515/TLR.2006.011.

Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second Language Research*, **19**(4), 329–358. https://doi.org/10.1191/0267658303sr226oa.

Adrada-Rafael, S. (2017). Processing the Spanish imperfect subjunctive: Depth of processing under different instructional conditions. *Applied Psycholinguistics*, **38**(2), 477–508. https://doi.org/10.1017/S0142716416000308.

Adrada-Rafael, S., & Filgueras-Gómez, M. (2019). Reactivity, language of think aloud protocol, and depth of processing in the processing of reformulated feedback. In R. Leow (Ed.), *The Routledge handbook of second language research in classroom learning* (pp. 201–213). Routledge.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59**(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bannard, C., & Lieven, E. (2009). Repetition and reuse in child language learning. In R. Corrigan, E. Moravcsik, H. Ouali, & K. Wheatley (Eds.), *Formulaic language: Volume 2: Acquisition, loss, psychological reality, and functional explanations* (pp. 299–321). John Benjamins.

Bardovi-Harlig, K. (2009). Conventional expressions as a pragmalinguistic resource: Recognition and production of conventional expressions in L2 pragmatics. *Language Learning*, **59**(4), 755–795. https://doi.org/10.1111/j.1467-9922.2009.00525.x.

Bardovi-Harlig, K., & Stringer, D. (2017). Unconventional expressions productive syntax in the L2 acquisition of formulaic language. *Second Language Research*, **33**(1), 61–90. https://doi.org/10.1177/0267658316641725.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. Available from arXiv:1506.0496

Boers, F., & Demecheleer, M. (2001). Measuring the impact of cross-cultural differences on learners' comprehension of imageable idioms. *ELT Journal*, **55**(3), 255–262. https://doi.org/10.1093/elt/55.3.255.

Boers, F., & Lindstromberg, S. (2012). Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics*, **32**, 83–110. https://doi.org/10.1017/S0267190512000050.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a Lexical approach to the test. *Language Teaching Research*, **10**(3), 245–261. https://doi.org/10.1191/1362168806lr195oa.

Bowles, M. A. (2008). Task type and reactivity of verbal reports in SLA: A first look at a L2 task other than reading. *Studies in Second Language Acquisition*, **30**(3), 359–387. https://doi.org/10.1017/S0272263108080492.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. Routledge.

Bowles, M. A., & Leow, R. P. (2005). Reactivity and type of verbal report in SLA research methodology: Expanding the scope of investigation. *Studies in Second Language Acquisition*, **27**(3), 415–440. https://doi.org/10.1017/S0272263105050187.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, **27**(6), 668–683. https://doi.org/10.1016/0749-596X(88)90014-9.

Carrol, G., & Conklin, K. (2014). Getting your wires crossed: Evidence for fast processing of L1 idioms in an L2. *Bilingualism: Language and Cognition*, **17**(4), 784–797. https://doi.org/10.1017/S1366728913000795.

Carrol, G., & Conklin, K. (2017). Cross language lexical priming extends to formulaic units: Evidence from eye-tracking suggests that this idea 'has legs'. *Bilingualism: Language and Cognition*, **20**(2), 299–317. https://doi.org/10.1017/S1366728915000103.

Carrol, G., & Conklin, K. (2020). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*, **63**(1), 95–122. https://doi.org/10.1177/0023830918823230.

Carrol, G., Conklin, K., & Gyllstad, H. (2016). Found in translation: The influence of the L1 on the reading of idioms in a L2. *Studies in Second Language Acquisition*, **38**(3), 403–443. https://doi.org/10.1017/S0272263115000492.

Carrol, G., Littlemore, J., & Dowens, M. G. (2018). Of false friends and familiar foes: Comparing native and non-native understanding of figurative phrases. *Lingua*, **204,** 21–44. https://doi.org/10.1016/j.lingua.2017.11.001.

Cieślicka, A. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research*, **22**(2), 115–144. https://doi.org/10.1191/0267658306sr263oa.

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

Cieślicka, A., & Heredia, R. (2011). Hemispheric asymmetries in processing L1 and L2 idioms: Effects of salience and context. *Brain and Language*, **116**(3), 136–150. https://doi.org/10.1016/j.bandl.2010.09.007.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, **29**(1), 72–89. https://doi.org/10.1093/applin/amm022.

Cooper, T. C. (1999). Processing of idioms by L2 learners of English. *TESOL Quarterly*, **33**(2), 233–262. https://doi.org/10.2307/3587719.

Coulmas, F. (1994). Formulaic language. In R. Asher (Ed.), *Encyclopedia of language and linguistics* (pp. 1292–1293). Pergamon Press.

Dai, Z., & Ding, Y. (2010). Effectiveness of text memorization in EFL learning of Chinese students. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 71–87). Continuum.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, **32**, 17–44. https://doi.org/10.1017/S0267190512000025.

Ellis, R. (1991). Grammatically judgments and second language acquisition. *Studies in Second Language Acquisition*, **13**(2), 161–186. https://doi.org/10.1017/S0272263100009931.

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language*, **6**(1), 22–42.

Gibbs, R. W., & O'Brien, J. E. (1990). Idioms and mental imagery: The metaphorical motivation for idiomatic meaning. *Cognition*, **36**(1), 35–68. https://doi.org/10.1016/0010-0277(90)90053-M.

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning*, **66**(2), 296–323. https://doi.org/10.1111/lang.12143.

Hubers, F., Cucchiarini, C., & Strik, H. (2020). Second language learner intuitions of idiom properties: What do they tell us about L2 idiom knowledge and acquisition? *Lingua*, **246**, 102940. https://doi.org/10.1016/j.lingua.2020.102940.

Inhoff, A. W., & Liu, W. (1997). The perceptual span during the reading of Chinese text. In H. Chen (Ed.), *Cognitive processing of Chinese and related Asian languages* (pp. 243–266). The Chinese University of Hong Kong Press.

Irujo, S. (1986). Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language. *TESOL Quarterly*, **20**(2), 287–304. https://doi.org/10.2307/3586545.

Jiang, N. (2011). *Conducting reaction time research in second language studies*. Routledge.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, **91**(3), 433–445. https://doi.org/10.1111/j.1540-4781.2007.00589.x.

Jiang, S., Jiang, X., & Siyanova-Chanturia, A. (2020). The processing of multiword expressions in children and adults: An eye-tracking study of Chinese. *Applied Psycholinguistics*, **41**(4), 1–31. https://doi.org/10.1017/s0142716420000296.

Kim, C. (2016). L2 learners' recognition of unfamiliar idioms composed of familiar words. *Language Awareness*, **25**(1–2), 89–109. https://doi.org/10.1080/09658416.2015.1122025.

Kim, H. R., & Bowles, M. (2019). How deeply do second language learners process written corrective feedback? Insights gained from think-alouds. *TESOL Quarterly*, **53**(4), 913–938. https://doi.org/10.1002/tesq.522.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest Package: Tests in Linear Mixed Effects Models (Version 2.0-36) [Computer software]. https://www.jstatsoft.org/v082/i13

Laufer, B. (1989). A factor of difficulty in vocabulary learning: Deceptive transparency. *AILA Review*, **6**(1), 10–20.

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Riebl, H., & Singmann, H. (2021). Estimated Marginal Means, aka Least-Squares Means (Version 1.6.3) [Computer software]. https://github.com/rvlenth/emmeans

Leow, R. P. (1993). To simplify or not to simplify: A look at intake. *Studies in Second Language Acquisition*, **15**(3), 333–355. https://doi.org/10.1017/s0272263100012146.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior. *Studies in Second Language Acquisition*, **22**(4), 557–584. https://doi.org/10.1017/s0272263100004046.

Leow, R. P., & Mercer, J. D. (2015). Depth of processing in L2 learning: Theory, research, and pedagogy. *Journal of Spanish Language Teaching*, **2**(1), 69–82. https://doi.org/10.1080/23247797.2015.1026644.

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, **26**(1), 35–57. https://doi.org/10.1017/s0272263104026129.

Leow, R. P., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research*, **30**(2), 111–127. https://doi.org/10.1177/0267658313511979.

Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, **61**(1), 30–44. https://doi.org/10.1006/brln.1997.1876.

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory and Cognition*, **36**(6), 1103–1121. https://doi.org/10.3758/MC.36.6.1103.

Lindstromberg, S., & Boers, F. (2008). The mnemonic effect of noticing alliteration in lexical chunks. *Applied Linguistics*, **29**(2), 200–222. https://doi.org/10.1093/applin/amn007.

Liontas, J. I. (2003). Killing two birds with one stone: Understanding Spanish VP idioms in and out of context. *Hispania*, **86**(2), 289–301. https://doi.org/10.2307/20062862.

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, **45**(2), 267–290. https://doi.org/10.5054/tq.2011.247708.

Myles, F., & Cordier, C. (2017). Formulaic sequence (FS) cannot be an umbrella term in SLA: Focusing on psycholinguistic FSs and their identification. *Studies in Second Language Acquisition*, **39**(1), 3–28. https://doi.org/10.1017/s027226311600036x.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Nekrasova, T. M. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning*, **59**(3), 647–686. https://doi.org/10.1111/j.1467-9922.2009.00520.x.

Park, J., Cardwell, R., & Yu, H. T. (2020). Specifying the random effect structure in linear mixed effect models for analyzing psycholinguistic data. *Methodology*, **16**(2), 92–111.

Pawley, A., & Syder, F. H. (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics*, **7**(5), 551–579. https://doi.org/10.1016/0378-2166(83)90081-4.

Perception Research Systems. (2007). *Paradigm stimulus presentation*. http://www.paradigmexperiments.com

R Development Core Team. (2021). *R: A Language and Environment for Statistical Computing (Version 4.1.0) [Computer software]*. Vienna, Austria : R Foundation for Statistical Computing.

Sandra, D. (1990). On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *The Quarterly Journal of Experimental Psychology Section A*, **42**(3), 529–567. https://doi.org/10.1080/14640749008401236.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, **18**(1), 55–88. https://doi.org/10.1177/026553220101800103.

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, **11**(2), 285–301. https://doi.org/10.1515/cllt-2014-0016.

Siyanova-Chanturia, A., & Janssen, N. (2018). Production of familiar phrases: Frequency effects in native speakers and second language learners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **44**(12), 2009–2018. https://doi.org/10.1037/xlm0000562.

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, **36**(5), 549–569. https://doi.org/10.1093/applin/amt054.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, **27**(2), 251–272. https://doi.org/10.1177/0267658310382068.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). 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**(3), 776–784. https://doi.org/10.1037/a0022531.

Spöttl, C., & McCarthy, M. (2004). Comparing knowledge of formulaic sequences across L1, L2, L3, and L4. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 191–226). John Benjamins.

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, **18**(5), 523–534. https://doi.org/10.1016/s0022-5371(79)90284-6.

https://doi.org/10.1017/S0142716421000552 Published online by Cambridge University Press

Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory and Cognition*, **37**(4), 529–540. https://doi.org/10.3758/MC.37.4.529.

Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbolic Activity*, **9**(4), 247–270. https://doi.org/10.1207/s15327868ms0904_1.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, **61**(2), 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 153–172). John Benjamins.

van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology*, **96**(1), 19–30. https://doi.org/10.1037/0022-0663.96.1.19.

Van Ginkel, W., & Dijkstra, T. (2020). The tug of war between an idiom's figurative and literal meanings: Evidence from native and bilingual speakers. *Bilingualism*, **23**(1), 131–147. https://doi.org/10.1017/S1366728918001219.

Vanlancker-Sidtis, D. (2003). Auditory recognition of idioms by native and nonnative speakers of English: It takes one to know one. *Applied Psycholinguistics*, **24**(1), 45–57. https://doi.org/10.1017/s0142716403000031.

Vilkaitė, L. (2016). Are nonadjacent collocations processed faster? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **42**(10), 1632–1642. https://doi.org/10.1037/xlm0000259.

Vilkaitė, L., & Schmitt, N. (2019). Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations? *Applied Linguistics*, **40**(2), 329–354. https://doi.org/10.1093/applin/amx030.

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, **16**(2), 180–205. https://doi.org/10.1093/applin/16.2.180.

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, **32**(4), 430–449. https://doi.org/10.1093/applin/amr011.

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition*, **35**(3), 451–482. https://doi.org/10.1017/S0272263113000107.

Wolter, B., & Yamashita, J. (2014). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, **36**(5), 1193–1221. https://doi.org/10.1017/s0142716414000113.

Wolter, B., & Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition*, **40**(2), 395–416. https://doi.org/10.1017/s0272263117000237.

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

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, **20**(1), 1–28. https://doi.org/10.1016/s0271-5309(99)00015-4.

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, **44**(4), 647–668. https://doi.org/10.5054/tq.2010.235998.

Zheng, H. (2019). *The processing of two types of Chinese idioms by L1 and L2 speakers*. (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana.

Cite this article: Zheng, H., Bowles, MA., and Packard, JL. (2022). NS and NNS processing of idioms and nonidiom formulaic sequences: What can reaction times and think-alouds tell us? *Applied Psycholinguistics* **43**, 363–388. https://doi.org/10.1017/S0142716421000552