THE PROCESSING OF TWO TYPES OF CHINESE IDIOMS BY L1 AND L2 SPEAKERS

BY

HANG ZHENG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in East Asian Languages and Cultures
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Jerome L. Packard, Chair
Associate Professor Melissa A. Bowles
Associate Professor Misumi Sadler
Associate Professor Chilin Shih

ABSTRACT

Chinese idioms originated from the pre-Qin dynasty (~ 221 A.D.) and are the linguistic and cultural heritage of the Chinese civilization. However, only 2.47% of the still-in-use idioms have a regular structure that follows the modern language rules (Wang & Wang, 2010). How are the irregular forms processed and understood? This is the core question that this dissertation project sets out to investigate.

A critical observation about idioms is that the figurative meaning of an idiom often has nothing to do with the literal meanings of its component words; therefore, how would Chinese speakers comprehend idioms? What about Chinese second-language learners? In Chinese, there are two major categories of idioms: 惯用语 *guan-yong-yu* (GYY) conventional-use-language, and 成语 *cheng-yu* (CY) fixed-language. GYYs are often used in informal and spoken contexts, and CYs in formal and written contexts (Chen & Chen, 1994). How do native speakers of Chinese process the two types of idioms? Would the two types of idioms be perceived and processed in different ways by native and nonnative speakers? Does this categorization of Chinese idioms have psycholinguistic grounds? This dissertation project sets out to address these questions through various measurements.

The quantitative data of speakers' metalinguistic judgments and response times reveal the processing patterns for the two types of idioms. First, native speakers process GYYs in the same way as they process the rule-generated phrases with the constituent words being accessed; while CYs are processed differently from their novel phrase counterparts with the internal words not being activated in a priming experiment. Secondly, both GYYs and CYs demonstrate processing advantages over their matched non-idiomatic formulaic sequences (FSs) during native speakers'

processing. Nonnative speakers, however, process shorter phrases faster than longer phrases regardless if they are idioms or FSs.

The qualitative data gathered from a think-aloud procedure reveal that even advanced learners of Chinese tend to analyze idioms, and the dichotomous judgment data (e.g., Yes-or-No judgments on grammaticality of the stimuli) tends to overestimate learners' knowledge of idioms.

The investigation of idiom processing in this dissertation presents comprehensive comparisons for the two types of Chinese idioms. The studies also contribute an idiom database that provides descriptive norms for further studies on idiom processing. Native speakers' ratings and different processing patterns observed for GYYs and CYs provide a psycholinguistic account to distinguish the two types of idioms. Chinese learners' thought processes contribute new evidence for modeling second language idiom processing.

To My Grandfather

ACKNOWLEDGMENTS

Over the past seven years of my studies at the University of Illinois, there have been many people without whom I would never have reached this point. First and foremost, I would like to thank my dissertation committee, who have helped me from breeding an idea and designing experiments to completing the dissertation. I am deeply indebted to my advisor, Professor Jerome Packard, who not only guided me through the journey of Chinese morphology and psycholinguistics, but also taught me how to think critically when I had questions, to be confident when I had self-doubts, and to be optimistic when I faced failures. It was his encouragement, guidance, and mentorship that helped me overcome many difficulties in my studies and that I will cherish forever.

I would like to thank Professor Melissa Bowles, who inspired me to use think-aloud protocols to study the topic of idiom processing. The L2 study in the dissertation developed out of a term paper in her class, Instructed Second Language Acquisition, and since then, she had devoted lots of time to helping me revise experiments, inspect statistical analyses, edit drafts, and prepare the manuscript. Even though we were from different departments, Professor Bowles always treated me as her own student and generously offered her resources when she found that I was in need. Most importantly, she always believed that my project was promising and would be successful. It was her faith in me that kept giving me the strength to proceed.

I would also like to express my gratitude to Professor Misumi Sadler who shared my research interest in formulaic sequences and generously shared her books with me and provided insights on this topic. Other than research, my teaching skills also benefited greatly from her class, East Asian Language Pedagogy. When I was applying for teaching-oriented positions, she spent her precious time reading and revising the statement of teaching philosophy. In all the time

I had been working with her, she showed great care not only to my academic accomplishments but also my career planning and personal welfare.

My thanks also go to Professor Chilin Shih, who helped me tirelessly to revise the experiment designs and to inspect the test materials. Whenever I was stuck in choosing between two designs or interpreting conflicting findings, her genius could always help me think of a solution, or point out a new perspective to see the problem. More importantly, her rigorous scholarship and scientific attitude toward research deeply infected me and will continue to inspire me in my career.

I would also like to extend my gratitude to Professor Zhang Bo at Beijing Language and Culture University, who advised my undergraduate study and my first master thesis, and had never stopped teaching, guiding, and caring ever since.

My colleagues and friends, You Li, Yihan Zhou, Junghwan Maeng, Xiaohui Zhang, Tianyu (Sophie) Qin, Wenqi Yang, Kailu Guan, Yuyun Lei, Jiani Lin, Zheshu Zhu, and Jaehee Park were the constant sources of my happiness and strengths. The moments of laughter and tears we shared together will be cherished in my memories.

Last but not least, I dedicate this dissertation to my grandfather, who had helped me make every important decision in my life, and whose kindness, bravery, rigor, and attention to detail motivated me to grow into the person I am now.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Idioms (e.g., 'kick the bucket') are fixed phrasal sequences perceived as highly conventional by native speakers. Idioms often have unified figurative meanings that are different from their literal interpretations, if there are any. For some idioms (e.g., 'by and large'), their overall meanings cannot be derived from the internal syntactic and semantic components, but for others (e.g., 'pop the question'), the figurative meanings are fairly transparent. Some idioms (e.g., 'throw in the towel') only make sense when people know their metaphorical meanings; others (e.g., 'let the cat out of the bag'), however, can be interpreted both literally and metaphorically. The heterogeneity of idioms has posed problems for the models of literal language processing, and has stimulated a considerable amount of research discussing the comprehension of idioms from both the speaker-internal (mental representation) and speaker-external (structural properties) perspectives. This dissertation sets out to explore the processing of idioms from the learner-internal point of view.

The research targets are two types of Chinese idioms namely, the four-character idioms (or *cheng-yu* fixed-language, henceforth CY), such as 一举两得 *yi-ju-liang-de* one-move-two-gains, "to gain two things by one move[1]," and the three-character idioms (or *guan-yong-yu* conventional-use-language, henceforth GYY), such as 走后门 *zou-hou-men* walk-back-door, "get in through the back door; pull strings." The two types of idioms have some universal characteristics that have been observed in idioms across different languages and also carry some Chinese language-specific features. This project aims to compare the processing of the two types

[1] All translations were initially taken from The Dictionary of Chinese Idioms (Shi, Wang & Zhang, 2006); for selected idioms not listed in the dictionary, translations were taken from two online translation portals (Google Translate and Baidu Translate).

of Chinese idioms with focus given to the lexical representation and lexical activation during NSs and nonnative speakers' (NNSs) processing. The general question that I want to ask is how the different linguistic properties of the two types of idioms are projected in NSs and NNSs' lexicons.

The dissertation has three goals: (a) identifing the linguistic properties of two types of Chinese idioms through NSs' ratings on multiple linguistic dimensions, (b) discussing the issue of the internal lexical activation in the visual processing of idioms, and (c) probing NNSs' processing strategies in idiom comprehension processing through the behavioral data and NNSs' concurrent verbal reports. In Study 1, I collected descriptive norms for 425 Chinese idioms. The distributions of the norms showed that NSs' intuitions about the lexical properties of the two types of idioms are distinct from one another. The high-frequency CYs were rated as more familiar, compositional, and formal than the high-frequency GYYs. This distinction was also observed in native speakers' on-line processing of the two idiom types. In Study 2, it was observed that GYYs were recognized in a similar fashion to rule-generated novel phrases while CYs were processed significantly differently from rule-generated novel phrases. These findings suggest that GYYs and CYs may have different lexical natures and lexical representations, and they challenge the claim that GYYs are more like single words and CYs are more like phrases (Zhou, 1998). In Study 3, I utilized qualitative data, response time, quantitative data, and think-aloud protocols to compare L1 and L2 speakers' processing of idioms and high-frequency novel phrases. The protocols showed that L2 learners used more analytical strategies to process idioms than L1 speakers did, which is basically consistent with the patterns revealed by the reaction time and dichotomous judgment data. Besides, L2 learners' protocols revealed that not all the idioms that had been correctly recognized in the dichotomous judgment task were fully understood. The

unique contribution of think-aloud protocols also highlighted the significance of introspective data in second language acquisition (SLA) research.

The dissertation is organized as follows. Chapter 2 reviews different hypotheses proposed for idiom processing and different measures used to examine idiom processing. Chapter 3 reports the descriptive norms of 425 Chinese idioms of two types. The results are discussed and compared with previous findings about Chinese and non-Chinese idioms. In Chapter 4, a priming study investigating the issue of lexical activation during the processing of the two types of Chinese idioms is presented. Chapter 5 presents a think-aloud study examining L1 and L2 speakers' processing of idioms in comparison with non-idiomatic lexical bundles. In Chapter 3 through 5, reviews of individual studies that are specifically related to the study are provided. Chapter 6 collectively discusses the findings of the three studies.

CHAPTER 2: BACKGROUND AND LITERATURE REVIEW

The Definitions of Idioms

The phenomenon of idioms has gained the attention of linguists and psycholinguists since the 1950s, when syntacticians (e.g., Chomsky, 1965) first noticed that idioms were different from literal language such that their compositions defy recursive grammar rules. Due to this observation, definitions of idioms mainly concentrate on describing the relationship between the idiom's entirety and its compositional elements. Weinreich (1969, cited from Bobrow & Bell, 1973) defines an idiom as "a complex expression whose meaning cannot be derived from the meanings of its elements." Chafe (1970) distinguishes two meanings of idioms, stating that the literal meaning is, in fact, able to be derived from the meanings of the words in the string, but it is the idiomatic meaning that is unable to be derived from the meanings of the individual words. Just as Swinney and Cutler (1979) and Schweigert (1986) illustrate, for the idiom "kick the bucket," its idiomatic meaning "die" has little to do with the meanings of either the verb "kick" or the noun "bucket," unlike its literal meaning, "to strike a pail with the foot." Gibbs (1980) points out that despite the non-compositional characteristics, the literal interpretation and the intended metaphorical referent may still be related to various degrees. Take the colloquial idiomatic expression "let the cat out of the bag," for example; the constituent element "let" and "out" have obvious associations with the idiomatic meaning "reveal the secret."

Because of the unique relationship between a whole idiom and its constituents, idiom comprehension challenges the standard view of language comprehension, according to which understanding a phrase entails recognizing the internal elements that compose the phrase based on some grammatical relations (Cacciari & Tabossi, 1988). It is also claimed that because idioms

have lost their metaphorical origins over time, current speakers directly stipulate the metaphorical meaning that an idiom is associated with in the mental lexicon (Chomsky, 1965; Cruse, 1986), and in order to learn the meanings of idioms, speakers need to remember the arbitrary links between idiom forms and their nonliteral meanings (Gibbs & O'Brien, 1990). Other researchers (e.g., Cutting & Bock, 1997) claimed that although not being directly related to the idiomatic meaning, the constituent words and the internal syntax still play an active role in idiom comprehension. Through manipulating different linguistic factors that could potentially impact how people would understand idioms, a series of hypotheses regarding idiom comprehension has been proposed based on empirical findings.

The Hypotheses of Idiom Comprehension

The idiom list hypothesis (ILH, Bobrow & Bell, 1973)

Searle (1968; cited from Ortony, Schallert, Reynolds & Antos, 1978) proposes that the processing of idioms entails three stages. At the first stage, the literal meaning of the word string is computed. Secondly, speakers check the literal interpretation against the context. Thirdly, if there is a conflict between the literal meaning and the context, speakers reinterpret the phrase idiomatically. The necessary condition of the idiomatic meaning retrieval is that a literal analysis has been attempted and failed. Bobrow and Bell (1973) take one step further, postulating that upon checking against the context and realizing the word string is not literal, a so-called "idiom mode" would be turned on, and the idiomatic meaning would be directly retrieved from an idiom list in the mental lexicon, which is separate from the single word lexicon. The proposal of "idiom mode" processing was based on the finding that when the test material includes a larger portion of idiomatic sentences, the number of first-perceived-as-idiom trials increased; when the

test materials include a larger portion of literal sentences, the number of first-perceived-as-idiom items decreased.

The lexical representation hypothesis (LRH, Swinney & Cutler, 1979)

Also supposing that idiomatic meaning is directly retrieved from the lexicon, the LRH, on the other hand postulates that people comprehend idioms in the same way as they process single words instead of storing idioms in a separate list. Although the literal meaning and the idiomatic meaning are activated simultaneously, the direct retrieval of the figurative meaning is much faster than the computation of the literal meaning by means of verbatim analysis. A faster and successful retrieval of the idiomatic meaning terminates the process of literal analysis unless the idiomatic interpretation was found to not fit the context. In that case, the literal analysis will continue. This hypothesis was proposed on the basis of Swinney and Cutler's findings that in making timed acceptability judgements for phrases, native speakers were faster and more accurately at making decisions on idioms than their novel phrase counterparts generated by replacing the first or last words of the idioms.

The direct access hypothesis (DAH, Gibbs, 1980)

The LRH was attacked by Gibbs (1980), who argued that the literal analysis is not an obligatory process in idiom comprehension based on the observation that the nonliteral meanings of idioms tend to be the first to come to mind. For example, "to die" is the first meaning of "kick the bucket" that would come to speakers' minds. Gibbs (1985, 1986) thereby proposed that people can directly access an idiom's figurative meaning bypassing the computation of the literal meanings unless figurative meaning is not proven to be conflicting to the context, and whether the literal or idiomatic interpretation was retrieved first is context-dependent.

Along with the proposal of the possibility that idiomatic meanings are directly accessed, Gibbs (1980) also suggests a continuum view of idiom processing. Popiei and McRae (1988) and Schweigert's (1991) analyses further provide a more detailed portrait of the concept of continuum, suggesting that idioms' properties vary along continuums of figurative meaning, familiarity, and literal plausibility; an idiom's position on these continuums determines whether its figurative or literal meaning is activated initially.

The configuration hypothesis (CH, Cacciari & Tabossi, 1988)

Although, the DAH seems to provide a straightforward solution that is particular to idiom comprehension, people cannot simply inhibit the automatic accessing of the constituent words because it is an automatic process that meanings of words that are attended to are activated (Stroop, 1935). Studies of on-line idiom processing have suggested that people cannot bypass the literal meanings of an idiom's constituents (Cacciari & Glucksberg, 1991; Cacciari & Tabossi, 1988; Connine & Blasko, 1993; d'Arcais & Giovanni, 1993; Tabossi & Zardon, 1993). How does one account for the quick retrieval of the idiomatic meaning without necessarily presuming the inhibition of automatic literal processing? Cacciari and Tabossi (1988) viewed idiom comprehension as an on-going process of configurating a string of words until sufficient input has rendered the configuration recognizable as an idiom. The individual words that participate in the configuration are the same lexical items that are accessed ordinarily during comprehension. For example, the word "take", as a lexical entry is activated in order to understand the regular phrase "take a book" as well as the idiom "take the bull by the horns." Cacciari and Tabossi further propose that there is a part of an idiom serving as the key of the string that determines when the idiom can be identified. For many English idioms, before the very last word, the string of words can be perfectly literal (e.g., 'let the cat out of ___(bag)').

The idiom decomposition hypothesis (IDH; Gibbs & Nayak, 1989)

Gibbs and his colleagues (Gibbs and Nayak, 1989; Gibbs, Nayak and Cutting, 1989; Gibbs, Nayak, Bolton and Keppel, 1989) introduced the IDH mainly for the purpose of calling attention to the lexical dimension of idioms, namely, decomposability, when conducting experiments on idiom comprehension because Gibbs and colleges found that idioms' syntactic and semantic flexibilities vary along this dimension and consequently affect about how the meaning of the constituents contribute to the overall meaning of an idiom (Gibbs & Nayak, 1989: p. 104). A decomposable idiom (e.g., 'break the ice') has a fairly transparent semantic structure in that the overall figurative meaning is highly derivable from the combination of the individual words. A non-decomposable idiom (e.g., by and large) cannot be understood by analyzing the internal syntactic and semantic components. This hypothesis adds another continuum-compositionality-to idioms' lexical properties, arguing that not all idioms are equally non-decomposable (see, for example, Gibbs and Nayak, 1989; Gibbs, Nayak and Cutting, 1989; Gibbs, Nayak, Bolton and Keppel, 1989; Titone and Connine, 1994).

The hybrid views (Cutting & Bock, 1997; Sprenger, Levelt & Kempen, 2006)

Different from previous hypotheses, the more recent models have a hybrid view which does not presuppose that idiom comprehension is so different from the comprehension of literal language. There are different models proposed from the hybrid perspective.

First, Cutting and Bock (1997) posit that in idiom comprehension, syntactic rules and the lexicon interact with each to activate the idiomatic meaning. The syntactic part of the model consists of a set of rules that create a structural frame. Take the verb-object idiom 'spill the beans' for instance. The syntax contributes the verb-object frame [V the O] with two specific syntactic slots. The lexicon consists of a network of nodes for linguistic units of various grain-

sizes, and contributes terminal notes, called lemma, (i.e., words, morphemes, phonemes, or concepts) that categorically fit the slots in the frame [V the O]. For example, the idiom "break the ice" is associated with three individual lemmas (break, the, and ice). However, for a less decomposable idiom like "kick the bucket," it will be linked to a lexical-concept lemma "die" which is also pre-listed in the lexicon. The authors claim that in idiom processing, the syntax and the literal meanings are both active and affect the processing based on the findings that when speakers were presented with two syntactically or semantically similar idioms (e.g., "hold your tongue" and "button your lip") at the same time, they tended to make more blend errors in the recall test (e.g., button your tongue).

On the basis of Cutting and Bock' (1997) hybrid model, Sprenger, Levelt, and Kempen (2006) go one step further, proposing that there is a *superlemma* layer that can mediate the activation of constituent words' literal meanings and retrieval of idioms' figurative meanings. The evidence was from the facilitative effects observed in a series of primed production tasks. When native speakers of Dutch were asked to recall, complete, or name some idiomatic expressions (e.g., . . .*viel buiten de boot* "to be excluded from something") and their matched novel phrases (. . .*ging met de boot* "took the boat") primed by a word that is included in both the idiomatic and literal phrases (e.g., *de boot* "boat"), more priming effects were found for idiom phrases than literal phrases. The findings jointly indicate that literal word meanings become active during idiom production.

As mounting evidence has shown that idiom comprehension is determined by multiple factors, the CH, IDH, and the hybrid views are favored over the earlier proposed models due to their broader explanatory powers.

Measures of Idiom Processing

This section reviews the research methods that are commonly used in the idiom processing literature and are theoretically relevant to the experimental designs of the studies in this dissertation. From the 1970s to the present, measurements of idiom processing and comprehension mainly fall into the following three types: metalinguistic judgments and ratings, response times, and elicited productions.

Metalinguistic judgments

Metalinguistic judgments are language speakers' intuitive statements, attitudes, and opinions on the language stimuli of interest. Speakers' judgments can be a dichotomous yes-or-no judgment or a 1–5 or 1–7 point scaled rating. The data are used to reflect the categorical or abstract nonverbalizable knowledge about a given language for psychometric purposes (Chraudron, 1983). Studies may employ metalinguistic ratings or judgments of native and non-native speakers to investigate specific research questions or to build a database to provide reliable psychometric variables for future research.

Metalinguistic judgments and ratings have long been employed to investigate the comprehension of idioms (Bobrow & Bell, 1973; Gibbs, 1980; Gibbs & Nayak, 1989; Gibbs, Nayak, Bolton, & Keppel, 1989; Sam, Glucksber, & Cacciari, 1994; Tabossi, Fanari, & Wolf, 2008). For example, Gibbs (1980) asked one group of subjects to read short stories in which the last sentences contain target idioms. After reading the last sentence, a paraphrase of the last sentence, providing either literal or idiomatic interpretation of the containing idiom, appeared. The task was to make a true-or-false judgment on whether the paraphrase of the idiom fits the context of the story, as quickly as possible. The control group of subjects were asked to judge the same paraphrases to the idiom-carrier sentences without reading a context story. Paraphrase

judgment errors showed that there is a strong bias to idiomatic interpretation over the literal interpretation without context; however, with context, subjects judged both the literal and idiomatic interpretations equally correctly. Based on these results, the author suggested that ease of comprehension for idioms may be more a matter of how conventional the contexts where idioms occur is than whether it is more literal or more metaphoric. Gibbs and Nayak (1989) used six idiom-paraphrase similarity judgment tasks examined why some idioms can be syntactically changed (e.g., "John laid down the law" can be passivized as "The law was laid down by John") and still retain their figurative meanings, while other idioms cannot be syntactically altered without losing their figurative meanings (e.g., "John kicked the bucket" cannot be passivized into "The bucket was kicked by John"). Based on the findings, the authors concluded that the decomposability of an idiom plays an important role in people's assumptions about the syntactic flexibility, and proposed the idiom decomposition hypothesis. Gibbs, Nayak, and Cutting (1989) used native speakers' ratings to investigate the relationship between decomposability and syntactic flexibility of idioms. In their study, subjects were presented with a list of idioms along with their figurative definitions. The task was to do a decomposability rating, to rate on a 1-7 scale, the degree to which the individual words made some semantic contribution to the figurative interpretations of syntactically frozen idioms versus syntactically flexible idioms. Results showed that about 60% of the syntactically flexible idioms were considered as decomposable idioms. However, only 10% of the syntactically frozen idioms were considered as non-decomposable ones. Thus, they concluded that the syntactic flexibility of an idiom may not impact speakers' assumptions about the semantic compositionality of the idiom.

Another line of research using metalinguistic judgment or ratings is norming research, whose primary goal is to collect norms that describe the intuitions of a large sample of

population on certain basic linguistic units (i.e., morphemes, words, and idioms) to provide psycholinguistic experiments with reliable statistical parameters (Noll, Scannell, & Craig, 1979). To our knowledge, the first norming study for English idioms was done by Popiei and McRae (1988). In their survey, ninety-six native speakers of English rated sixty English idioms on a 1-7 scale on three dimensions, familiarity, assessing how well a speaker understands an idiom, literality, assessing the likelihood an idiom's literal interpretation is realizable in the real world, subject frequency, assessing how often a speaker encounters an idiom. The largest norming study for English idioms was conducted by Bulkes and Tanner (2017), which collected descriptive norms for 870 common idioms from 2,100 subjects through online surveys.

Study 1 in this dissertation is a norming study of 425 Chinese idioms. The descriptive norms collected in Study 1 contribute experimental materials and variables for Study 2 and 3 in this dissertation. Table 3.1 in Chapter 3 displays a list of previously published norming studies for idioms in different languages.

Response times (RTs)

Response time, also referred to as reaction time or response latency, is the amount of time individuals take in responding to a stimulus while performing a task. Jiang's work (2013) is devoted to the issue of how to design and implement empirical research where reaction time is the primary data. The premise of using RT data, as Jiang points out, is that the cognitive effort that individuals make to process language under particular conditions can be inferred by observing how long it takes them to respond. RT data is often measured in online tasks because it quantifies processing efforts while language comprehension is on-going, or "during its operation" (Swinney, 1979: 647). Because of the requirement of measuring this narrow window of time, a language-related task for this purpose cannot be too complicated and must be time-

sensitive. For that reason, grammaticality/acceptability judgment tasks (GJT/AJT) and the priming paradigm are ideal experimental methods to elicit time-sensitive data.

GJT/AJT was adapted from a word recognition task devised by Meyer and Schvaneveldt in the 1970s (Meyer & Schvaneveldt, 1971; Meyer, Schvaneveldt, & Ruddy, 1974). The procedure has been used in many psychology and psycholinguistics experiments ever since. In a typical GJT/AJT, subjects are presented with a set of linguistic stimuli about which they must judge the acceptability/grammaticality by indicating "Yes" or "No." The elicited responses provide two types of data: judgment accuracy and RTs (refer to Tremblay, 2005 for a methodological review). To maximize the chance of estimating the "on-going" processing moment before it's completed or at least immediately after its completion, participants are usually instructed to respond as quickly as possible. There are two major advantages of GJT/AJT (Nation, 1990, 2001, 2013; Laufer, 1997; Meara & Milton, 2003). First, it is possible to include a large number of items in one test because the tasks are easy to design and conduct, and data are easy to collect and analyze. Second, because of its brief format, test takers are not likely to lose concentration during the short test sessions.

Research on idiom processing and comprehension has long employed this method (e.g., Swinney & Culter, 1979; Gibbs, Nayak & Cutting, 1989; Burt, 1992; Glass, 1983; Mueller & Gibbs, 1986; Tabossi, Fanari & Wolf, 2009). For example, Gibbs, Nayak, and Cutting (1989) asked subjects to judge the acceptability of some idioms with various degrees of decomposability and found that subjects were significantly faster at judging the decomposable idioms than non-decomposable idioms. The authors' explanatinon was that the processing time advantage for decomposable idioms was due to that people attempt to do compositional analysis in idiom comprehension. For decomposable idioms, people can assign independent meanings to its

individual words and are able to quickly recognize how the meaningful parts constitute the figurative meaning. For non-decomposable idioms, it would be more difficult to determine the oveall meaning through individual parts, and thus people need longer verification time to process the idioms. The authors also considered these findings as evidence for the IDH. Burt (1992) conducted three experiments, letting subjects make speeded acceptability judgements about idiomatic, literal-everyday phrases and nonsense phrases. In the first experiment subjects responded faster to idioms and everyday phrases than their control phrases. Idioms with high metaphoric transparency were also processed faster than the ones with low metaphoric transparency. The interpolation of the constituent words in phrases did not abolish the processing advantage of idioms over control phrases. However, idioms with high-frequency initial words were processed as rapidly as idioms with low-frequency initial words. Phrases with different length, on the other hand, were responded to differently. The findings did not support the LRH, which proposes that the idioms are stored as lexical entries and are recognizable upon the encounter of the initial word. Tabossi, Fanari, and Wolf (2009) used a semantic judgment paradigm, where subjects read a sentence and pressed a button as soon as the sentence became meaningful to them. The results showed that participants were faster at judging decomposable idioms, non-decomposable idioms, and everyday clichés than at judging their matched novel phrases. The authors claimed that the results provided evidence for the CH, suggesting that idioms are processed faster because they are familiar and recognizable rather than they are semantically frozen.

In the SLA literature, phrase GJT/AJTs in visual and auditory modalities have been widely adopted to study NNSs' processing of multi-word lexical units (e.g., Jiang, 2000; Jiang & Forster, 2001; Jiang, 2002). The experiment design usually pairs up an idiom and a matched

novel phrase (non-idiom) with similar frequency and orthographic complexity and compares which type of stimuli is responded to faster and more accurately. However, GJT/AJT has proven to be problematic in assessing L2 learners' knowledge because the dichotomous judgment (Yes or No) elicited a considerable amount of random guesswork even with ungrammatical/unacceptable fillers added. The guesswork may undermine the validity of the findings. Shillaw (1999) found that Japanese learners tend to have a more conservative strategy, giving more NO answers, which yielded an under-estimation of their vocabulary knowledge. For Belgian students on the contrary, Boers, Eyckmans, and Stengers (2007) found more YES responses were provided, which yielded an over-estimation of their knowledge. R Ellis (1991) asked Chinese learners of English to think aloud when redoing a GJT which the participants did one week before and found learners were inconsistent in 22.5% in their judgments. Targeting the 'random guesswork' problem, N. Schmitt, D. Schmitt, and Clapham (2001) revised Nation (1990)'s Vocabulary Levels Test by using a word-meaning association task, in which learners must choose a correct meaning for a word from a list of options. Other researchers use additional approaches, such as think-aloud protocols, to complement the reaction times and the on-line judgments to establish "an adequate level of internal validity" (Leow, Grey, Marijuan & Moorman, 2014).

Study 3 of this dissertation uses GJTs with a think-aloud procedure to examine idiom processing by both L1 and L2 speakers. Both RT data and dichotomous judgment accuracy are gathered and analyzed in comparison with the think-aloud verbalizations.

Priming paradigms are another time sensitive concurrent data collection method that are commonly used in the research on idiom processing. Priming is a psychological technique whereby subjects are exposed to a pair of stimuli in which the perception of one stimulus

influences the response to a subsequent stimulus, usually without the subject's conscious knowledge of the influences. The rationale is that there are priming effects that would trigger subjects to retrieve an item from memory when primed by an associated concept (e.g., Anderson, 1983; Collins & Loftus, 1975; Quillian, 1967). The more association between the prime and target, the more facilitative effects would be observed between the prime and target. The amount of the association can be mapped into the RTs and response accuracy to produce priming effects (e.g., Ratcliff, 1978; Ratcliff & McKoon, 1978). For example, when the word "cat" is primed by a semantically related word, "dog," it is easier to be decided as a word, and the latencies to recognize "cat" is shorter than if the prime word is "six" (Forster & Davis, 1984). Because priming effects can be very subliminal and instantaneous, subjects' reactions on the targets are sensitive to the stimulus onset asynchrony (SOA—the amount of time between the onset of the prime and the target). SOAs were often determined in the pilot phase of the experiments or based on the previously published norms. In literature, it has been reported that for a word-to-word prime-target design, the SOA should be below 250 ms, to prevent subjects from detecting the purpose of the experiment and consequently developing a strategy to perform the task (Neely, 1976, 1977; McRae & Boisvert, 1998).

In the literature of idiom processing, priming effects are adopted to examine the speakers' perception in regard to the relationship between idiom's overall meaning and its constituent words (e.g., Cacciari & Tabossi, 1988; Cutting & Bock, 1997; Sprenger, Levelt, and Kempen, 2006; Holsinger, 2013; Titone & Connine, 1994, 2014; Cacciari & Corradini, 2015). For example, Cacciari and Tabossi (1988) conducted three cross-modal primed lexical decision tasks to examine the lexical access in the processing of idioms with low and high predictability. Results showed that when the idioms were predictable, subjects were faster at recognizing the

targets that were idiomatically related to the prime idioms than literally related targets; when idioms cannot be recognized as idioms until encountering the last word, subjects were faster on the targets literally related to the last words of prime idioms; when target words were presented 300 ms after the end of prime idiom was auditorily presented, subjects were equally fast at literally related and idiomatically related target words. Titone and Libben (2014) used three cross-modal priming experiments to investigate how linguistic differences among idioms in semantic decomposability, familiarity, and literal plausibility affect figurative meaning activation with the prime idiom-carrier sentences auditorily presented and target words being visually presented. The authors found that the figurative meaning activation increased as the idiom string unfolds to 1000 ms later. Second, different lexical factors such as familiarity, compositionality, and literality plausibility of idioms modulate the figurative activation at different time points. These results strongly contradict the IDH and suggest that idiom retrieval is constrained by multiple linguistic factors in a time-dependent fashion.

In Study 3 of this dissertation, priming paradigms are used to select experimental materials and examine the processing of two types of Chinese idioms. Chapter 4 will present a more detailed review of previous idiom research using priming paradigms.

Experimental production

Experimental production methods vary along a continuum, from loosely structured production such as think-aloud protocols and elicited production, where participants essentially have the freedom to say whatever they want to say, albeit with the prompts or hints from experimenters; to highly structured production such as naming, primed recall, and elicited imitation, where participants are asked to repeat a particular language string (Ambridge & Rowland, 2013).

Both loosely and highly constrained production methods have been employed in idiom processing studies (e.g., Gibbs & O'Brien, 1990; Cacciari & Glucksberg, 1995) and have generated some important processing models (e.g., Cutting & Bock, 1997). A highly structured production example is Cutting and Bock's (1997) cued recall experiment. In the study, subjects were presented with a pair of idioms with similar or different syntactic and semantic elements, and after a brief moment were asked to recall one of the two idioms. The purpose of this recall procedure was to elicit recall errors. The elicited errors showed that subjects tended to mix the two idioms (blend errors) in their recall when the two idioms share the same syntactic structure or contain the same constituent word. Based on the finding, the author proposed the hybrid processing, supposing that syntactic frame and lexical concept are both activated during the process of idiom comprehension. A loosely structured production example is Gibbs and O'Brien's (1990) study which elicited imagery protocols. In three experiments, speakers were asked to form and describe mental images while they read some idioms. Next, subjects were asked a series of prompt questions about the events related to their mental images (What caused the action? Was the action done Intentionally? In what manner is the action done? What was the consequence of the action? What would have happened if the action had not been done?). The researchers found high consistency in subjects' images of idioms' metaphorical meanings despite differences in their surface forms (e.g., "spill the beans" and "let the cat out of the bag" both mean "reveal the secret") while the imagery protocols associated with the controlled literal phrases were quite varied. Based on the regularity in people's images, the authors concluded that idioms are not "dead" metaphors. Conflicting findings were obtained by Cacciari and Gluscksberg (1995) using similar mental-image production task. The authors suspected that because some idioms (often referred to as ambiguous idioms) can convey both a literal and an

idiomatic meaning, if mental images consistently reflect the idiomatic meaning, people must be able to ignore literal meaning in some way. Therefore, they conducted experiments to investigate the potential interference between literal and idiomatic meaning and found that the images obtained for ambiguous idioms overwhelmingly associated with the literal meanings of the idioms rather than with their idiomatic meanings. In the discussions of both studies, the authors did mention a potential problem of the method that the instruction of forming a mental image or the prompt questions could manipulate subjects' thought processes in a biased way and cause the conflicting findings.

In the literature on idiom processing by L2 learners, loosely controlled productions are very frequently used both qualitatively and quantitatively. Abel (2003) examined the subjects' thoughts qualitatively in an exit survey after they rated their familiarity with some idioms. Subjects were asked what they did when they encountered an unknown idiom in the experiment. The majority answered that they considered the literal meaning of the individual words and tried to put together an idiomatic meaning that made sense. Based on the subjects' utterances, the author concluded that NNSs actually "decompose" idioms. NNSs' interpretations of idioms were elicited through a think-aloud procedure by Cooper (1999), used as both qualitative and quantitative data. In the study, participants were asked to verbalize their thoughts as they think of the meaning of some idioms. Analysis revealed that most of the participants employed a variety of strategies that are quite different from how L1 speakers comprehend idioms.

In this dissertation, Study 3 employs think-aloud protocols in conjunction with GJTs to investigate L2 learners' processing strategies of Chinese idioms and non-idiomatic phrases.

Think aloud (TA) protocols

The think-aloud procedure has been widely used by SLA researchers to gather data about

learners' thought processes or explicit knowledge for various theoretical and applied purposes. A typical TA procedure starts with one or two sentences, which briefly reiterate why the participants are being asked to think aloud but without giving away any information about the goal of the study. As an example demonstrated by Bowles (2008): "In this experiment, I am interested in what you think about when you complete these tasks. In order to find out, I am going to ask you to THINK ALOUD as you work through the mazes." Following the brief instruction of the rationale of the task, a more specific instruction on how to think-aloud while performing a task must be provided, including thinking-aloud involves exactly what they are expected to do, what language they should use in verbalizing their thoughts, how detailed their think-aloud is required to be (cf. Bowles, 2010). After the participants indicate that they understand the instructions, some practice trials are usually conducted to familiarize participants with either the think-aloud procedure or the task they will performing. Finally, in the real trial, to ensure validity, one researcher should be present with participants and remind them to think aloud whenever they fall back to silent thinking.

Bowles (2010) provides an overview of research using TA data and a meta-analysis of research, which finds this procedure to be valid if implemented appropriately. TA is viewed as a means of discovering the depth of processing, the amount of attention and strategies employed by learners when processing L2 input (e.g., Leow, Hsieh, and Moreno, 2008; Morgan-Short, Heil, Botero Moriarty & Ebert, 2012; Leow, Grey, Marijuan & Moorman, 2014).

As a versatile data collection tool, TA is employed in studies of either qualitative or quantitative or the combination of both methods. Mackey, Gass, and McDonough (2000) and Leow, Grey, Marijuan and Moorman (2014) suggest that TA protocols can be used to complement the limitation of other concurrent data collection procedures. There is a good deal of

of research comparing think-aloud approaches to other data gathering techniques or language assessment measures. The purposes are either to validate or question a particular research methodology or to get information about the participants' thought processes, strategies or depth of knowledge that single measures may not suffice or that other instruments may not be able to provide. For example, Kamimoto (2005; cited from Milton, 2010) introduced think-aloud protocols in an lexical decision task (LDT), in which, learners must think aloud by speaking out whatever on their minds when they make a yes-or-no lexical judgment. Godfroid and Schmidtke (2013) triangulate think-aloud protocols, eye-tracking, and pretests and posttests to investigate incidental vocabulary learning. Rebuschat, Hamrick, Riestenberg, Sachs, and Ziegler (2015) through triangulating think-aloud protocols, retrospective verbal reports, and other subjective assessments such as confidence ratings, investigated how NNSs allocate awareness under different incidental learning conditions. To investigate ESL learners' depth of vocabulary knowledge and lexical inferencing strategies, Nassaji (2006) compared learners' reading comprehension through contrasting a TA reading condition and a silent reading condition. Rosa and O'Neill (1999) utilize think-aloud protocols along with a recognition task to investigate how the learner's intake may be affected by the allocation of awareness and by different levels of explicit presentation. Morgan-Short, Heil, Botero-Moriarty, and Ebert (2012) compare think-aloud reading and traditional silent reading comprehension to discuss whether attending to grammatical or lexical form while reading for meaning would affect the comprehension of the text.

The TA procedure is also an ideal tool to study " individuals with varying levels of expertise within a certain domain of knowledge" (Fonteyn, Kuipers & Grobe, 1993). Therefore, researchers exploit the TA procedure to compare the depth of processing between L1 and L2

speakers. For example, in other to test the competing hypotheses regarding reading comprehension, Davis and Bistodeau (1993) compared the reading comprehension models through native language readers' TA verbalizations when they approached reading in their L1 and L2. The significant finding was that when reading in L1, participants used more bottom-up strategies (i.e., commenting on intrasentential features, focusing on individual words, or providing restatements) in comprehending low-frequency linguistic forms, but when reading in L2, novice L2 speakers tended to use more top-down strategies (i.e., predicting what was coming next, confirmation of the prediction, reference to antecedent information, or using encyclopedia knowledge to relate the information in the text).

One methodological controversy of think-aloud protocols that TA research must address is reactivity. Reactivity is the potential that the act of thinking aloud may alter subjects' cognitive processes while performing a task. However, both a meta-analysis study (Bowles, 2010) and empirical research (Leow & Morgan-Short, 2004; Bowles & Leow, 2005) have shown that the TA procedure does not have detrimental or facilitative effects on most of the SLA tasks with which it has been used. The authors also suggested that any research employing TA procedure should include a control group who does not think-aloud to ensure no detrimental effect of the act of thinking-aloud on subjects' performances.

The TA approach is employed in Study 3 in this dissertation. Participants are asked to think aloud while performing a GJT. The dichotomous judgment measures and TA productions are compared.

Research Targets

The research targets of this dissertation include two types of Chinese idioms, CYs and GYYs. CYs have a uniform surface structure and consist of four Chinese characters, most of which have traceable historical origins of various genres, such as mythology, classical allusion, or literature. Therefore, CY is more often used in formal or written language. GYYs come in varying numbers of characters (three characters about 98 percent of the time) and are conventional, figurative sayings that emerged from people's life experiences and were passed down orally. Thus, GYYs occur more often in informal or spoken language. Based on the formats and the figurative origins, CYs are more frozen and less analyzable than GYYs.

In spite of these differences, both types possess the characteristics that are universally observed in Idioms. One global observation is that the meanings of idioms, to different extents, cannot be derived directly from the meanings of their components. For example, the Chinese idiom 一举两得 *yi-ju-liang-de* one-move-two-gains "achieve two things at one stroke" is composed of two juxtaposed noun phrases, to get whose meaning, one simply needs to add a logical connective. By contrast, another nominal idiom 杯弓蛇影 *bei-gong-she-ying* cup-bow-snake-shadow "to mistake the shadow of a bow projected in one's cup as the shadow of a snake, indicating a false alarm or self-created suspicion" probably requires some prior knowledge of the figurative genesis behind the idiom to understand it. The same holds true for GYYs: the idiomatic meaning of 走后门 *zou-hou-men* walk-back-door "get in through the back door; pull strings" is more relevant to its component words than that of 翘辫子 *qiao-bian-zi* rise-braid-SUF[2] "to be dead". As a consequence, the standard models of language processing, which typically rely on the integration of syntactic and semantic information by certain rules cannot adequately

[2] SUF: suffix

account for the comprehension of all idioms. Another global characteristic of idioms also observed in Chinese is that they are syntactically defective such that syntactic operations are not always allowed on these phrases. For example, according to Chinese syntax, many verb-object phrases can undertake verb-object inverting by adding an aspectual ending to convert an active voice into a passive one, such as 开门 *kai-men* open-door "open the door" into 门开着 *men-kai-zhe* door-open-ASP[3] ("the door is open"). However, in the verb-object GYY 翘辫子 *qiao-bian-zi* rise-braid-SUF "to be dead", undergoes the verb-object inverting operation, the outcome 辫子翘着 *bian-zi-qiao-zhe* braid-rise-ASP "the braid has risen" will completely lose its figurative meaning. For other verb-object GYYs, syntactic manipulation would not cause the figurative meaning. For example, GYY 戴高帽 *dai-gao-mao* wear-tall-hat "to flatter someone" if undergoing passivation and changed into 被戴了高帽 bei-dai-le-*gao-mao* PASS[4]-wear-ASP-tall-hat "to be flattered", would maintain its idiomatic interpretations. However, as for CYs, the format is frozen and forbids any form of lexical or syntactic alternation. For example, if adding an omittable modifier marker 的 *de* to the CY 花花世界 *hua-hua-shi-jie* colorful-colorful-world-boundary "the dazzling human world with its myriad temptations", the idiom variant 花花的世界 *hua-hua-de-shi-jie* colorful-colorful-MOD[5]-world- boundary will lose its idiomhood. In fact, the Chinese term for CYs literally means "fixed language". Thus, when the fixedness of a CY is broken, it could be no longer regarded as an idiom.

In this dissertation, the processing of GYYs and CYs by L1 and L2 speakers is compared in three studies. Particularly in each study, the following questions will be addressed.

[3] ASP: aspectual marker
[4] PASS: passivation marker
[5] MOD: modifier marker

Research Questions

Study 1 reports descriptive norms for 182 GYYs and 243 CYs. All idioms are commonly used in Chinese and selected from the top frequency band from Google 1-gram database. In the study, descriptive norms are elicited from more than 2000 native speakers for the following dimensions: familiarity, meaningfulness, literality, compositionality, final-word predictability, and linguistic register. The variables are selected on the basis of (1) the universality in the literature (2) the theoretical relevance to the experimental investigations of the current study.

The second study investigates the relationship between constituent words of an idiom and the whole idiom form during the processing of idioms by native speakers. The key debate in the idiom processing hypotheses is that whether the individual words would be automatically accessed during idiom processing. Study 2 attempts to address this question using evidence of Chinese idiom processing. Particularly, I want to ask (1) if the internal word will be semantically activated during the early stage idiom processing of idioms, and (2) if there is internal semantic activation, would that be observed for both types of Chinese idioms?

Study 3 investigates how L2 learners comprehend idioms and non-idiomatic FSs. The central question is whether the processing of idioms in L2 is the same as or different from that in L1. To answer this question, three types of data, namely, RT, two-way metalinguistic judgments, and think-aloud protocols. The goal of triangulating the three data sources is twofold. One is to gain a comprehensive understanding of idiom acquisition in L2, and the other is to see if the three data will reveal the same or different patterns concerning L2 learners' processing and knowledge of Chinese idioms. To address these issues, the following research questions are asked:

1. What do RT data reveal about L1 and L2 speakers' knowledge of idioms and FSs?

2. What do dichotomous judgments reveal about L1 and L2 speakers' knowledge of idioms and FSs?

(a) Do L1 and L2 speakers judge the same stimuli in the two GJTs with the same degree of accuracy?

(b) Do L1 and L2 speakers judge the same stimuli consistently in the two GJTs?

3. What do TA verbalizations reveal about L1 and L2 speakers' knowledge of idioms and FSs?

(a) Can dichotomous judgments reflect L1 and L2 speakers' actual knowledge of idioms and FSs?

(b) Do L1 and L2 speakers use the same strategies to process idioms and FSs?

(c) Is the processing strategy correlated with the accuracy of dichotomous judgments?

CHAPTER 3: STUDY 1
DESCRIPTIVE NORMS FOR TWO TYPES OF CHINESE IDIOMS

This study details descriptive norms for 425 of the most commonly used Chinese idioms. The idioms consist of two types: three-character idioms, otherwise referred to as *guan-yong-yu* (GYY) 'conventional-use-language,' and four-character idioms, often referred to as *cheng-yu* (CY) "fixed-language." The first of two goals of this study is to compare the lexical and syntactic differences between GYYs and CYs. The second is to provide publicly available norms on the processing of Chinese idioms for future research. In order to facilitate opportunities for consensus and comparable findings on the idiom processing of other languages, five widely adopted linguistic dimensions of measurement from previous norming studies were used for this study: familiarity, meaningfulness, compositionality, literality, and last-word predictability. In addition, the language register that distinguishes the two types of Chinese idioms was also normed. The 425 chosen idioms were rated for the aforementioned dimensions by 2748 Chinese native speakers. Statistical analyses of the ratings and responses are discussed in comparison with previous findings.

Introduction

Sam Glucksberg (1993) cited a proverb "people who live in glass houses should not throw stones" to illustrate the relationship between the figurative meaning of an idiomatic expression and its constituent words. In this case, "glass houses" is a metaphor for "vulnerability," and "throw stones" is a metaphor for "criticize." As Glucksberg explained, the figurative meaning of the proverb can be derived from the allusions of the two metaphors. The

interpretation of this proverb raised several core questions about idiom comprehension, which are also essential to this study: (1) to what extent can the figurative meaning of an idiom be inferred from its internal components? (2) Can anyone who is a fluent speaker (e.g., a fluent English learner) understand the figurative meaning of an idiom? (3) Setting aside the figurative meaning, does the literal meaning of an idiom still make sense? (4) For an idiomatic expression, such as the proverb mentioned above, if a constituent (e.g., "throw stones") is missing, can speakers still associate the remaining part of the sentence with that particular proverb? Each of these questions relates to a dimension of idioms that plays a role in idiom processing and comprehension.

The first question—the relationship between constituents and overall meaning—involves the compositionality of an idiom. **Compositionality**, also referred to as decomposability, describes a feature of literal language that the meaning of a phrase can be obtained by combining the component words of the phrase according to the grammar of the language. Some grammarians (e.g., Chomsky, 1980) consider idioms a typical case of non-compositionality due to their defective grammar. A more widely adopted view was proposed by Nunberg, Sag, and Wasow (1994), who suggested that compositionality, in the context of idioms, refers to whether the constituents of an idiom contain any "identifiable parts of the idiomatic meaning" (p. 496). Some idioms (e.g., break the ice) may be more decomposable because the component words are connected to the overall figurative meaning in a literal way. Other idioms (e.g., kick the bucket) are less decomposable, which is the case when the literal meanings of the component words have no connection to the overall figurative meaning; this type of idiom is considered a stereotypical idiom. Compositionality is also a central issue in idiom processing. For a more decomposable idiom, it is assumed that its components are retrieved from the mental lexicon separately and

combined online through syntactic operations (Tabossi, Wolf, & Koterle, 2009). By contrast, a less decomposable idiom is assumed to be retrieved directly from the lexicon, and the comprehension will be disrupted if any internal components were altered (e.g., "kick the bucket" changed to "kick the pail"; Gibbs, Nayak, Bolton, & Keppel, 1989). Accordingly, the comprehension of decomposable and non-decomposable idioms is also realized through different processes.

The second question—whether the literal meaning of an idiom is capable of fitting a truth condition in the real world (Lakoff, 1986)—relates to the literality of idioms. **Literality**, also referred to as literal well-formedness, literal likelihood, literalness, or literal plausibility, indicates the likelihood that an idiom is used literally. The defining feature that distinguishes idioms from novel phrases is the figurative meaning that idioms carry (Nordmann, Cleland & Bull, 2014). However, in some cases, an idiom can also be used literally regardless of its figurative meaning (e.g., at the end of the day). The literality of an idiom is often a focus of studies on idiom processing. The research questions of such studies often involve, for example, whether the literal meaning and the figurative meaning of an idiom are initiated simultaneously (Swinney & Cutler, 1979), whether the literal meaning of the component words will be activated when processing a non-decomposable idiom (Titone & Connine, 1994), or whether speakers will respond to a less literal idiom faster than a more literal one (Swinney & Cutler, 1979; Mueller & Gibbs, 1987; Cronk & Schweigert, 1992).

The third question—whether prior experience or knowledge is needed to correctly understand the figurative meaning of an idiom—involves the familiarity and meaningfulness of an idiom. **Familiarity**, also referred to as subjective frequency, indicates how often a language user may encounter an idiomatic expression (Schweigert, 1986; Titone & Connine, 1994). The

familiarity of an idiom may influence how it is processed because less familiar idioms, although also are recognizable as idioms, are harder to retrieve than the familiar idioms (Schweigert, 1986). As for an unfamiliar idiom, it may even be recognized by a person as an idiom. For example, many idioms that are fairly familiar to native speakers of a given language may not be identified as idioms by even advanced second language learners of this language.

Meaningfulness represents a speaker's confidence about how well he/she knows what an idiom actually means (Libben & Titone, 2008). These two factors are the most commonly measured dimensions in existing norming studies for idioms (see Table 3.1 for reference) and have been shown to be important influences on idiom recognition and comprehension (Titone & Connine, 1994) as well as metaphor processing (Blasko & Connine, 1993).

The fourth question—whether an idiom fragment can trigger language users to retrieve the remaining part of the idiom—concerns an idiom's predictability. **Predictability** refers to the likelihood of an idiom being completed in a fill-in-the-blank (cloze) task. Generally, the missing part in these tests is at the end of the idiom. Cacciari and Tabossi (1988) proposed that the predictability of the final word may elicit the dominant activation of the idiomatic interpretation. In other words, whether or not the final word is predictable is associated with whether the idiom is understood through the ongoing analyses of the sequence or whether the analysis takes place instead through direct retrieval prior to phrase offset (Libben & Titone, 2008).

Thus far, I have mentioned five dimensions that have been frequently used to define and distinguish different types of idioms. They are also the variables that previous norming studies for idioms frequently examined. Table 3.1 displays such studies for idioms in English and non-English languages.

Table 3.1: Norming studies for English and non-English idioms

	English idioms					Non-English idioms			
Study	Bulkes & Tanner (2017)	Libben & Titone (2008)	Titone & Connine (1994)	Schweiger t & Cronk (1992)	Popiei & McRae (1988)	Li, Zhang, & Wang (2016)	Citron, Cacciari, Kucharski, Beck, Conrad, & Jacobs (2016)	Bonin, Meot, & Bugaiska (2013)	Tabossi, Arduino, & Fanari (2011)
Language	English	English	English	English	English	Chinese	German	French	Italian
No. of idioms	870	210	171	390	60	350	619	305	245
No. of subjects	2,100	160	226	164	96	735	249	187	740
No. of ratings/item	≥100	≥30	≥28	≥30	≥40	≥20	≥30	≥23	≥40
Rating scale	5-point	5-point	7-point	5-point	7-point	7-point	7-point	5-point	7-point
Survey format	Online	Booklet	Booklet	Booklet	Booklet	Booklet	Online	Booklet	Booklet
Dimensions	FAM, MEA, COM, LIT, PRE	FAM, MEA, Global COM, Local COM, LIT, PRE	FAM, COM, LIT, PRE	FAM, COM, LIT, PRE	FAM, LIT, Sf	KNO, FAM, Sf, COM, LIT, PRT, AoA	KNO, FAM, CON, FIG, ST, EV, Arousal	KNO, FAM, Sf, COM, LIT, PRE, AoA	KNO, FAM, COM, LIT, PRE, AoA, SF
Selection criteria	• Idiom dictionary • Modernity • Sentence compatibility	• Idiom dictionary • Syntax	• Idiom dictionary	• Idiom dictionary	• Previous empirical studies	• Idiom dictionary • Familiarity ratings	• Idiom websites • Syntax	• Idiom websites • Syntax	• General dictionary • Syntax • Popularity

For all of the tables and figures in this chapter, FAM= familiarity; MEA= meaningfulness; LIT= literality; COM= compositionality; PRE= predictability. KNO= knowledge; Sf= subject frequency; ST= semantic transparency; EV= emotional valence; AoA= age of acquisition; SF= syntactic flexibility; CON= concreteness; FIG= figurativeness.

Unlike other languages, the Chinese language categorizes idioms based on a pragmatic dimension, linguistic register. Although, linguistic register has not been particularly investigated as a dimension of idioms in the literature, it is not a new dimension of language. Ferguson (1994; p. 16) defined linguistic register as to refer to "the linguistic differences that correlate with different occasions of use." The term "register" is used to describe a wide variety of language forms that are determined by the context of use. Different lexical choice is often the most immediate index of different registers. One dimension of linguistic register is discourse modality—spoken discourse versus written discourse (Ravid & Berman, 2009), or otherwise referred to as colloquial language versus formal language (Clackson, 2010). Nayak and Gibbs (1990) first examined people's intuitions on the context appropriateness of idioms. The authors observed that although idiom "do a slow burn" and "flip one's lid" have similar idiomatic interpretation "get angry," they convey a subtle but significantly different concept so that it is appropriate to use them in different contexts. The authors also found that not only do people have a clear intuition on the discourse difference, but also the coherence between idioms and contexts facilitated the processing speed of idioms.

In the Chinese language, two major classes of idiomatic expressions, GYYs and CYS, emerge from discrimination based on discourse modality. CYs (e.g., 杯弓蛇影 cup-bow-snake-shadow "mistake the shadow of a bow projected in the water of one's cup as a snake - a false alarm") are rooted in ancient Chinese, a large portion of which originated from literary works, historical events, or legendary stories with traceable sources and quotations in classical texts. According to Xiao's (1987) statistical analysis of the sources of more than 4,000 CYs, 63%

originated from texts from the pre-Qin[6] Dynasty (~ 221 A.D.), 15% from the Wei, Jin, Southern, and Northern Dynasties (220~589 B.C.), 9% from the Sui and Tang Dynasties (581~907 B.C.), 6% from the Song Dynasty (960~1279 B.C.), and 2% from the Yuan and Ming Dynasties (1206~1644 B.C.). Therefore, the vocabulary of CYs is formal and different from that of contemporary Chinese, and the structure does not always abide by modern grammar rules. CYs have a uniformly fixed four-character format that is not subject to any change, substitution, or reduction. In a sentence, a CY is used as a whole and may take the subject, object, or predicate slot. GYYs (e.g., 敲竹杠 knock-bamboo-lever "take advantage of someone's being in a weak position") are conventional expressions that entail metaphorical meanings. They mainly stem from traditional customs, religious practices, and real historical events. Because GYYs are essentially a folk language created and passed down orally by people based on their life and work experience, they carry the characteristics of spoken language, such as simple structure and visualizable wording. Because GYYs are a type of colloquial language, most of them carry a strong emotional valence, such as a rhetorical tone or a derogatory implication (Wen, 1989, 2007). Therefore, it is important to know if a GYY is a complementary term or derogatory term in order to use it felicitously. Most GYYs have a fixed format of three characters but with more flexible structure. Lexical adding, removing, and replacing are sometimes allowed. Syntactic operations, such as insertion, dislocation, or passivation, are also applicable. Generally, a GYY is a verb-object phrase, a modification structure, or a verb-compliment structure, all of which can be found in modern Chinese. Therefore, it is more comprehensible to current Chinese speakers if used in a specific context.

[6] For example, some idioms (e.g., 哀鸿遍野 moaning-swan-spread all over-wild "a land swarming with famished refugees") originated from the *Classic of Poetry* (诗经).

Despite the difference in linguistic register, GYYs and CYs are idiomatic expressions and possess the general characteristics of idioms that have been observed in other languages. These global characteristics taken together with register may jointly influence the processing of Chinese idioms. However, to our knowledge, little empirical research has been conducted focusing on comparing GYYs and CYs. One goal of this study was to fill this research gap by comparing the lexical and syntactic properties of these two types of idioms. To do so, through a large-scale online survey, we collected descriptive norms for 425 high-frequency idioms based on six dimensions: familiarity, meaningfulness, compositionality, literality, predictability, and language register. The other goal of this study was to provide publicly available norms for future research on the comprehension and processing of these two types of Chinese idioms. The study essentially follows the protocols that have been widely adopted in previous norming studies listed in Table 3.1 and 3.2. Statistical analyses are presented and discussed in comparison with the findings of these studies.

Method

Participants

The participants for this study were recruited from four universities in China (n=2748). Electronic questionnaires were programmed via wenjuanxing.com and then distributed through the smart phone application WeChat by the students' class advisors in the departments from which the participants were recruited. To avoid one participant rating the same idiom for more than one dimension, participants' IP addresses were limited to allow a given IP address to log into only one of the study's questionnaires. To ensure all participants spoke standard Mandarin Chinese, participants were asked to self-report their Mandarin Chinese level ("standard,"

"standard with light accent," or "not standard"). Participants who reported their Mandarin Chinese as "not standard" were excluded from further analyses. All participants provided online informed consent before they preceded to one of the questionnaires and received a small cash payment for their participation.

Materials

A total of 425 idioms were selected following a three-step procedure. First, we manually extracted all GYYs and CYs listed in the *Contemporary Chinese Dictionary* (6^{th} edition, 2015; hereto forth referred to as the Dictionary). An effort was made to be as exhaustive as possible. This procedure yielded a total of 5881 items consisting of 2098 GYYs and 3783 CYs. Second, the two lists were run through the Google 1-gram database. Based on the raw token frequency, the top-ranked 300 items in each list were further extracted. Finally, based on the observations in the pilot phase, some of the top-ranked idioms are unfamiliar to native college students. Therefore, following Bulkes and Tanner's (2017) approach, the idiom pool was further narrowed to include only 182 GYYs and 243 CYs based on whether an idiom is commonly and widely used in the modern Chinese language. Our CY list had 16 items that coincided with Li et al.'s (2016) list.

Procedure

The 425 selected idioms were pseudo-randomly divided into eight lists, each containing 53 or 54 items with approximately equal number of GYYs and CYs. Each list was duplicated six time. Each duplication was developed into a questionnaires, testing one of six dimensions: familiarity, meaningfulness, compositionality, literality, predictability, and linguistic register. Each participant was only permitted to provide responses for one questionnaire via the WeChat smartphone application. Because each IP address was only allowed to log into one questionnaire

for this study, no participant rated one idiom for more than one dimension. Finally, a minimum of 50 unique ratings were elicited for each idiom on one dimension. After eliminating the invalid data (participants who self-reported being not standard in Chinese), we finally randomly selected an even number of 50 ratings from the valid data for each idiom on each dimension. For the dimensions of familiarity, meaningfulness, compositionality, and literality, participants were asked to rate on a 5-point Likert scale (1= lowest; 5 = highest). For the dimension of predictability, participants were given a list of idiom fragments with the last character missing. Their task was to type in a word to make the phrase grammatical and meaningful. The final rating measure was the proportion of participants who completed the phrase with an answer that made it the expected idiom. After rating the items, participants were asked to complete a brief language background survey by checking the corresponding boxes prior to exiting the questionnaire. Instructions for the dimensions were adapted from Libben and Titone (2008) and Bulkes and Tanner (2017) and provided in Chinese (see Appendix A). To ensure participants fully understood the task they were asked to complete, participants were allowed to contact the experimenter through WeChat (the smartphone application that was used to distribute the questionnaires) when they had any questions.

Dependent variables

Familiarity refers to how often a speaker may encounter an idiom based on their personal experience. For this dimension, participants were asked to "use a 1–5 scale to indicate how often you read, hear, or use the expressions in the questionnaire. Point 1 = never heard, read, or produced; point 5 = heard, read, or produced very often."

Meaningfulness assesses how much speakers think they know about the figurative meaning of an idiom. For this task, participants were asked to "rate the idioms on a scale of 1 to

5, depending on how well you know the figurative, non-literal meaning of the idiom. Point 1 = you have absolutely no idea what the idiom means. Point 5 = you are 100% certain of the idiom's meaning and could explicitly explain the meaning in your own words."

Compositionality measures to what extent the literal meanings of the constituent characters of an idiom are related to its overall figurative meaning. For this task, we adopted Bonin, Méot, and Bugaiska (2013) approach by providing participants with the figurative definition along with each item. The definition of each idiom was copied from the *Dictionary*. For example, 一石二鸟 *yi-shi-er-niao* 'one-stone-two-bird': "一个举动达到两个目的 to achieve two goals with one move." Then, participants were asked to "use a 1–5 scale to rate whether the idiom is decomposable. 'Decomposable' means if its constituent parts contribute to the meaning of the expression. Point 1 = absolutely not decomposable; point 5 = completely decomposable."

Literality assesses the possibility that an idiom is used literally in the real world. In the task, participants were asked to "use a 1–5 scale to judge to what degree you find the expression to have plausible literal meaning in any context. Point 1 = absolutely not plausible; point 5 = completely plausible."

Predictability refers to the likelihood that speakers will complete an idiom fragment idiomatically. In this study, the missing part of an idiom is the last character. Participants were asked to "read the incomplete phrases and type in the blank the first words coming to your mind; make the phrase grammatical and meaningful. For each idiom fragment (e.g., 打官___, *da-guan*____, 'play-bureaucratic___') , the final rating measure is the proportion of participants who completed the sentence idiomatically (e.g., 腔 *qiang* 'accent').

Linguistic register evaluates the pragmatics of language, for example, the capacity of a linguistic form to index culturally recognizable activities or the context of language use (Agha,

2004). The present study specifically evaluates which language register, written (formal) or spoken (informal), the idioms belong to. Participants were asked to "choose which linguistic register you think these expressions belong to, written (formal) language or spoken (informal) language." A written register was assigned point 1 and a spoken register was assigned point 0. The higher an idiom is scored, the more likely the idiom indexes as formal language.

Syntactic structure is an experimenter-coded variable. In Li, Zhang, and Wang's study, CYs were coded into seven different syntactic structures: VO (verb-object), SM (structure of modification), SV (subject-verb), VV (verb-verb), VOVO (double VO), SMSM (double SM), and SVSV (double SV). This categorization was used in the present study. As for GYYs, we coded four structures[7]: VO (verb-object), SM (structure of modification), VC (verb-compliment), and SV (subject-verb). Table 3.2 presents example idioms for syntactic structure.

Table 3.2: Examples of idioms with different syntactic structures

	Structure	Example	Literal translations and glosses
CY	VO	不择手段	*Bu-ze-shou-duan* Not-choose-hand-method "Use unscrupulous divisive tactics"
	SM	花花世界	*Hua-hua-shi-jie* Colorful-colorful-world-boundary "The dazzling human world with its myriad temptations"
	SV	热血沸腾	*Re-xue-fei-teng* Hot-blood-boiling-rise "Burning with righteous indignation"
	VV	未雨绸缪	*Wei-yu-chou-mou* Not-rain-silk-pretend "Make provision in good times for bad days"
	VOVO	谈天说地	*Tan-tian-shuo-di* Talk-sky-speak-earth "Talk of everything under the sun"
	SMSM	五花八门	*Wu-hua-ba-men* Five-pattern-eight-category "Of a wide variety"
	SVSV	日新月异	*Ri-xin-yue-yi* Day-new-moon-different "Alter from day to day"
GYY	VO	开夜车	*Kai-ye-che* Drive-night-car "Stay up all night working"
	SM	铁公鸡	*Tie-gong-ji* Iron-male-chicken "A stingy person"
	SV	黑吃黑	*Hei-chi-hei* Black-eat-black "One illegal party bullies the other by coercive means"
	VC	恨不得	*Hen-bu-de* Hate-not-get "Be eager to (achieve something)"

[7] Because Li et al.' study only investigated CYs, the categorization of GYYs was not based on Li et al.'s study.

To summarize, in this study, six continuous variables are ratings (on familiarity, meaningfulness, compositionality, literality) and scores (on predictability and linguistic register) gathered from NSs' surveys. One categorical variable (syntactic structure) is coded by the experimenter.

Results and Discussion

The norms

The descriptive norms for each of the 425 idioms are presented in Appendix B (for GYYs) and Appendix C (for CYs). Table 3.3 presents the descriptive statistics of responses for GYYs and CYs for the six dimensions.

Table 3.3: Descriptive statistics for the norms (N=50) of each dimension by idiom type

		FAM	MEA	COM	LIT	PRE	REG
GYY	Mean	4.329	4.683	3.541	3.618	0.709	0.429
	St. Deviation	0.496	0.270	0.519	0.474	0.286	0.194
	Skewness	-1.566	-2.256	-0.148	-0.559	-0.814	0.185
CY	Mean	4.652	4.808	4.109	3.949	0.944	0.819
	St. Deviation	0.154	0.112	0.415	0.387	0.104	0.118
	Skewness	-1.126	-1.020	-0.777	-0.247	-2.241	-1.134

In all tables and figures in this chapter, REG refers to register.

The data for predictability indicates the proportion of participants completing the phrases as the target idioms; the data for register indicates the possibility the idiom occurs in the formal context. The data on most dimensions were not normally distributed. Familiarity and meaningfulness for both GYYs and CYs were strongly negatively skewed. Given that all idioms were in the top frequency band in the corpus, it is not surprising that they are all familiar to native speakers. Compositionality and literality were slightly negatively skewed for both types of idioms. Predictability was slightly negatively skewed for GYYs and strongly negatively skewed

for CYs, which indicates that CYs are more predictable than GYYs. Register was slightly positively skewed for GYYs and strongly negatively skewed for CYs, indicating a greater portion of GYYs were considered to be informal expressions and most CYs were considered to be formal expressions. Given that GYYs originate from folk language and the genesis of CYs is classical Chinese language, the difference in register is predictable.

To further examine on which dimension the two types of idioms are similar or different, independent t-tests were computed with the average ratings on the items as dependent variables. Table 3.4 presents the results.

Table 3.4: Independent t-test results for ratings/scores of GYYs and CYs on six dimensions

	FAM	**MEA**	**COM**	**LIT**	**PRE**	**REG**
Mean of GYY	4.33	4.68	3.54	3.62	0.71	0.43
Mean of CY	4.65	4.81	4.11	3.95	0.94	0.82
t value	-8.48	-5.86	-12.12	-7.69	-10.56	-23.90
df	207	228	338	343	217	280
p value	0.00	0.00	0.00	0.00	0.00	0.00

As can be seen from Table 3.5 that GYYs and CYs were rated/scored significantly different on every dimension. The visual presentations of the rating/scoring difference between the two types of idioms are displayed in Figure 3.1 using box-whisker plots.

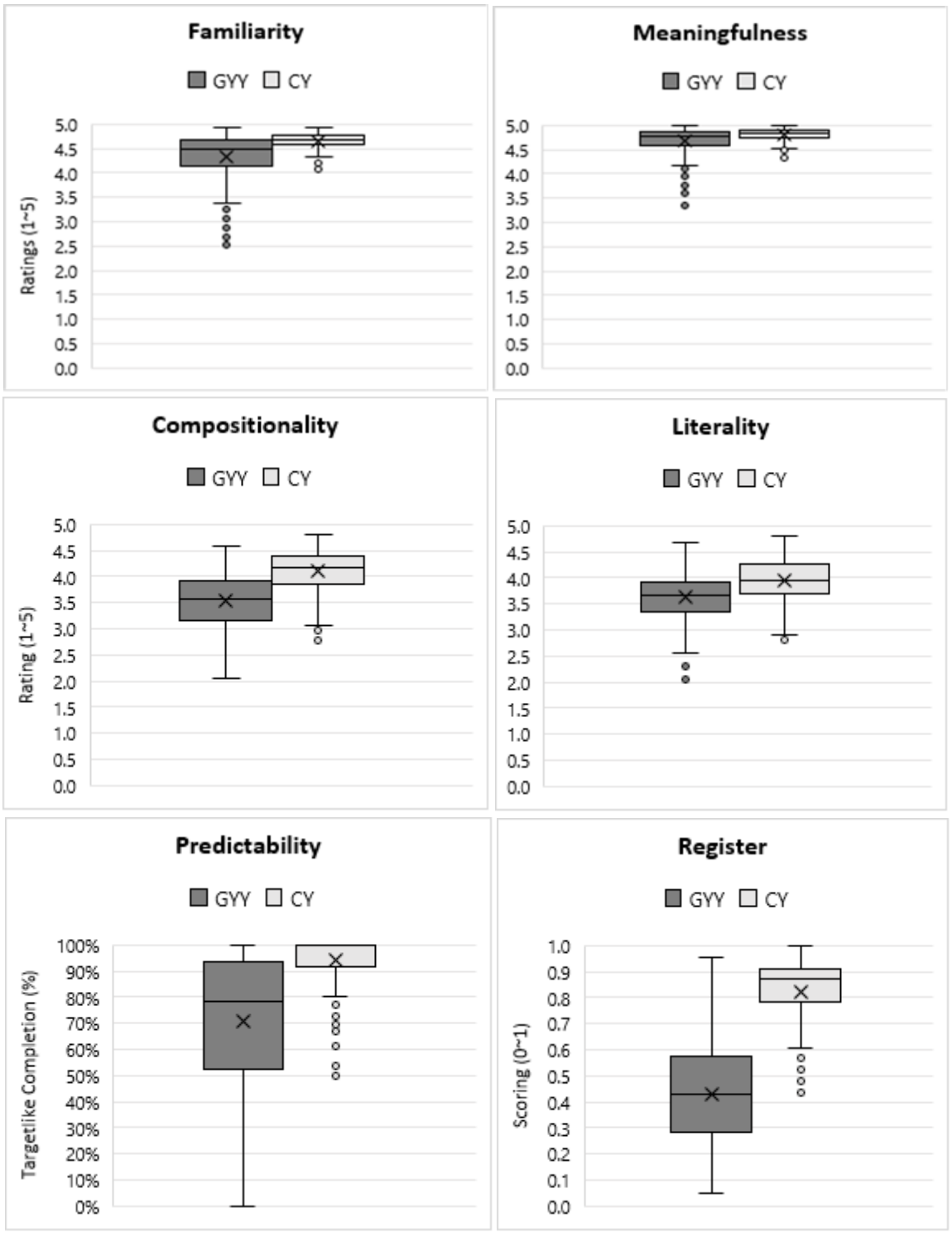

Figure 3.1: Box-whisker plots of average responses for idioms for the six dimensions by idiom type

The box-whisker plots in Figure 3.1 are visual representations of the distributions of the norms for the two types of idioms. The boxes represent the middle 50% of the scores, and the whiskers represent the upper and lower extremes. As can be seen in all six charts, the GYY boxes are lower than the CY boxes, indicating GYYs are rated lower than CYs across all dimensions. The results of independent t-tests also showed that the ratings/scores for the two types of idioms were significantly different on every dimension (see Table 3.5). The GYY whiskers also cross a wider range than the CY whiskers, indicating that the selected GYYs have greater internal variation than the selected CYs.

Overall, the scores for GYYs are generally lower than those for CYs. For meaningfulness, although the rating difference is statically different ($p<.000$), with the mid-point of the average ratings for both types greater than 4.5, indicates that participants had a high degree of understanding of the selected idioms, which further suggests that the participants' responses for the subsequent dimensions were based on knowledge, rather than on conjecture. For familiarity, GYYs (Mean=4.33) were rated significantly lower ($p<.000$) than CYs (Mean=4.65), and the range of the ratings for GYYs was also greater than that for CYs. This result suggests that although both the GYYs and the CYs were the most frequent of their own type, CYs are more frequently encountered in general than GYYs are. This difference may be caused by the different emotional valence of the two types of idioms. Emotional valence describes the emotional orientation—positive, negative, or neutral—that a linguistic form conveys (Russell, 1991; Citron et al., 2016). Most of the selected CYs are neutral expressions, while most GYYs are derogatory terms, only applicable in the context of sarcasm and ridicule. Due to the limited context, GYYs are encountered less often than CYs. The context difference is also reflected in the ratings of linguistic register, a dimension that is rated most differently between GYYs and

CYs. GYYs are rated to be more colloquial than formal, and CYs are rated more formal than colloquial. For compositionality, GYYs (Mean=3.54) were rated significantly lower ($p<.000$) than CYs (Mean=4.11), indicating that the overall meaning of the CYs is more closely related to the meaning of the component words than the overall meaning of the GYYs is. This result suggests that all or some of the individual words in a CY make semantic contributions to the understanding of that CY. For GYYs, however, the constituents "cannot be mapped individually in a one-to-one fashion" to the idiom's meaning (Glucksberg, 1994). Rather, the concept of the whole GYY extends to a new domain (Cacciari, 1994). The inferential path from the literal domain to the metaphor domain involves cultural conventions and is not always predictable. For literality, the ratings were also significantly lower ($p<.000$) for GYYs (Mean=3.62) than for CYs (Mean=3.69). Lower literality could mean a lower likelihood of the idiom fitting in a real-world situation, or it could mean the idiom is not directly meaningful if it is not borrowing from other domains of thought or experience (Lakoff, 1986; Cacciari, 1994). Given that the metaphor in GYYs usually involves semantic extension in another domain, lower literality indicates fewer GYYs than CYs make sense if interpreted literally. Predictability is one of the dimensions in which GYYs (Mean=0.71) and CYs (Mean=0.94) differ the most ($p<.000$) from the box plots. This result is not surprising, given the different degrees of syntactic flexibility. CYs are frozen forms. Therefore, when seeing the first three words of a CY, people can directly match the fragment with the pattern stored in their memories. In this case, the "idiom principle" applies: a slot is filled with a word which is a part of a prefabricated form (Erman & Warren, 2000). By contrast, GYYs are nor frozen forms. In this case, the "open choice principle" applies: a slot can be filled by any word abiding by the grammar rules. Tables 3.5 and 3.6 present the extreme cases for the two types of idioms.

Table 3.5: Example CYs rated as extremely low/high and the least predictable CYs

	Low	Meaning	Score	High	Meaning	Score
Extremely high/low cases and scores						
FAM	必由之路	the route one must take.	4.06	理所当然	taken for granted.	4.88
	叹为观止	take one's breath away in astonishment	4.07	见义勇为	act bravely for a just cause.	4.91
	舞文弄墨	show off literary skill.	4.15	游山玩水	make a sightseeing tour.	4.91
	雷霆万钧	as powerful as a thunderbolt.	4.20	再接再厉	make persistent efforts.	4.93
	天人合一	theory that man is an integral part of nature.	4.20	欢天喜地	be filled with great joy.	4.93
MEA	必由之路	the route one must take.	4.32	一模一样	be exactly alike	4.96
	引人入胜	lead one into the interesting part of something.	4.43	自言自语	keep on talking though no one is listening	4.96
	乐此不疲	delight in a thing and never get tired of it.	4.47	蠢蠢欲动	be eager for action	4.98
	不可或缺	absolutely necessary	4.49	多愁善感	always melancholy and moody	4.98
	有的放矢	have a definite object in view.	4.52	欢天喜地	be filled with great joy.	4.98
LIT	扑朔迷离	whirling; confusing the eye.	2.80	自言自语	keep on chattering though no one is listening.	4.64
	天马行空	a powerful and unconstrained style.	2.90	不知不觉	imperceptibly; unconsciously.	4.66
	迫在眉睫	extremely urgent.	3.07	一模一样	be exactly alike.	4.67
	炙手可热	the supreme arrogance and great power.	3.10	各式各样	every kind of .	4.72
	雷霆万钧	as powerful as a thunderbolt.	3.18	自始至终	from first to last.	4.81
COM	歇斯底里	hysteric.	2.77	大同小异	be the same in essentials but differ in minor points.	4.76
	扑朔迷离	whirling; confusing the eye.	2.80	一目了然	be apprehended at a glance.	4.76
	无可厚非	no ground for blame	2.88	赏心悦目	be pleasant to the eye	4.78
	青梅竹马	a friendship established in childhood.	2.97	冰天雪地	a frozen and snow-covered land.	4.78
	风花雪月	romantic themes.	3.06	知己知彼	know oneself and know the enemy.	4.80
REG	胡说八道	talk nonsense/rubbish.	0.30	出类拔萃	rise above the common herd.	1.00
	不知不觉	Unconsciously.	0.41	取而代之	replace someone's position.	1.00
	讨价还价	bargain with someone for a better deal.	0.41	得天独厚	be richly endowed by nature.	1.00
	一心一意	put one's whole heart into.	0.41	淋漓尽致	thoroughly; most incisive.	1.00
	心中有数	know the score.	0.41	别具一格	have a distinctive style.	1.00
Least predictable cases and scores						
PRE	不正之风	unhealthy tendency; bad working styles.	0.50			
	不知所云	do not know what others are talking about.	0.50			
	有的放矢	have a definite object in view.	0.54			
	不可收拾	unmanageable; irremediable.	0.55			
	一丝不挂	be in the nude.	0.56			

Table 3.6: Example GYYs rated as extremely low/high and the least predictable GYYs

	Extremely high/low cases and scores					
	Low	Meaning	Score	High	Meaning	Score
FAM	赶浪头	follow the trend.	2.52	不得了	terrible.	4.88
	面面观	multi-dimension view.	2.53	过日子	live a... Life.	4.90
	全武行	gang fight.	2.68	来不及	there's no time.	4.90
	打棍子	persecute.	2.89	靠得住	reliable.	4.93
	执牛耳	occupy a leading position.	2.96	做生意	do business.	4.93
MEA	软脚蟹	wuss.	3.35	绕圈子	make a detour; not straightforward.	4.95
	打板子	harshly criticize and punish.	3.61	来不及	there's no time.	4.95
	爬格子	writing hardly.	3.61	开绿灯	give free rein.	4.96
	开倒车	turn back the wheel of history.	3.66	做生意	do business.	4.96
	回马枪	back thrust.	3.76	打官腔	speak in a bureaucratic tone.	4.98
LIT	打秋风	seek gratuitous financial help.	2.05	来不及	there's no time.	4.49
	执牛耳	occupy a leading position.	2.31	翻白眼	feel angry/disappointed/embarrassed.	4.51
	泥饭碗	unstable job.	2.37	靠不住	unreliable.	4.51
	爬格子	writing hardly.	2.55	半辈子	half a lifetime	4.62
	全武行	gang fight.	2.60	看热闹	watch the scene of bustle.	4.68
COM	打板子	harshly criticize and punish.	2.04	挡箭牌	take someone or something as excuse.	4.47
	爬格子	writing hardly.	2.40	打哑谜	make puzzling remarks.	4.52
	执牛耳	occupy a leading position.	2.49	进一步	go a step further.	4.56
	二百五	a stupid person.	2.50	断头台	scaffold; guillotine.	4.57
	敲竹杠	take advantage of one's weak position to overcharge him.	2.57	等不及	can't wait to do.	4.57
REG	大不了	if the worst comes to the worst.	0.05	全武行	gang fight	0.82
	怪不得	no wonder .	0.05	空对空	empty and unrealistic.	0.82
	好意思	have the nerve (rhetorical).	0.10	莫须有	fabricated; unwanted.	0.84
	不得了	terrible; horrible; desperately serious.	0.10	集大成	epitomize.	0.85
	打交道	come into contact with be keen on face-saving.	0.10	执牛耳	occupy a leading position.	0.95
	Least predictable cases and scores					
PRE	赶浪头	follow the trend.	0.00			
	看得起	think highly of.	0.07			
	出人命	a death-causing accident.	0.07			
	差不离	more or less.	0.08			
	三不知	know nothing.	0.08			

Correlations

Following Bulkes and Tanner's (2017) approach, pairwise Spearman's rho correlations were calculated between the ratings/scores of the participants' responses to the six dimensions for GYYs and CYs, respectively. The results are displayed in Table 3.7 and Table 3.8.

Table 3.7: Correlation matrix with Spearman's rho calculations on seven dimensions for GYYs

	FAM	MEA	COM	LIT	PRE	REG
FAM	1.000					
MEA	.589**	1.000				
COM	.478**	.481**	1.000			
LIT	.583**	.462**	.578**	1.000		
PRE	.310**	.265**	.131	.265**	1.000	
REG	-.710**	-.492**	-.278**	-.482**	-.332**	1.000

**$p < 0.01$

Table 3.8: Correlation matrix with Spearman's rho calculations on seven dimensions for CYs

	FAM	MEA	COM	LIT	PRE	REG
FAM	1.000					
MEA	.321**	1.000				
COM	.448**	.380**	1.000			
LIT	.333**	.389**	.629**	1.000		
PRE	.250**	.050	.010	.139*	1.000	
REG	-.151*	-.048	-.143*	-.229**	-0.004	1.000

**$p < 0.01$ *$p < 0.05$

As can be seen from Table 3.7 and 3.8, positive correlations were observed between most of the dimensions except for the correlations between register and other dimensions. Linguistic register is negatively correlated with familiarity (rho=-.71) and meaningfulness (rho=-.492) for GYYs, indicating that the more formal a GYY is, the less frequently it is encountered and the lower the likelihood that it is known to native speakers. This result confirms that most GYYs are supposed to be used in an informal context and further suggests that native speakers may acquire GYYs by hearing them used in everyday situations. Register is slightly correlated with familiarity but not significantly correlated with meaningfulness for CYs, indicating that whether a CY is more formal or more colloquial has little to do how well people understand its meaning.

Given that most CYs were rated as formal language, we further speculate that this result may be related to the fact that most CYs are learned through formal schooling or reading. Therefore, formal or colloquial, the meanings of CYs were learned and remembered.

The ratings for literality were strongly positively correlated with those for compositionality for both GYYs and CYs. This result replicates Li, Wang, and Zhang (2016).'s norming study for Chinese idioms. In Libben and Titone (2008) and Bonin et al. (2013), negative correlations were found between literality and compositionality. In Bulkes and Tanner (2017), no correlation between literality and compositionality was found. Li, Wang, and Zhang (2016) attributed the different findings for Chinese compared to other languages to the nature of the written Chinese language, arguing that the meaning of an expression is visually more closely related to a pictographic language than that to an alphabetic language. Another reason for such a discrepancy, in our opinion, may be related to the constituent words of the idioms in the two types of languages. Chinese idioms are very condensed and compact structures. In most cases, the constituent words are all content words, each representing a concrete concept. By contrast, in other languages, function words often occur in idioms. All constituent words being content words increases the likelihood that some individual words are related to the figurative meanings and that some individual words' concepts are realizable in the actual words, for example, CY 舞文弄墨 *wu-wen-nong-mo* dance-character-play-ink "play with one's literary skill." Among the four constituent words, except the first word 舞 *wu* "to dance," the other three words are more or less related to the figurative meaning, and the verb-object structure 弄墨 *nong-mo* play-ink "play with ink" is also realistically plausible, such as imagining a calligraphist preparing ink. Although the CY is not fully literally plausible, due to the partially literal nature, this CY was rated 3.784 for compositionality and 3.426 for literality. Bulkes and Tanner also argued that language change

plays a role in causing these different findings. Given that there is a 20-year interval between their study and other previous norming studies, it is likely that some idioms have become more opaque over that course of time.

The ratings for familiarity were correlated with the scores for predictability (the proportion of participants who completed the idiom fragments as they were expected) for both CYs and GYYs. suggesting that the more familiar participants are with an idiom, the more likely they are able to complete it idiomatically. This finding is consistent with that of most previous norming studies (Li, Wang & Zhang, 2016.; Bulkes & Tanner, 2017; Titone & Connine, 1994). This finding supports the hybrid view of idiom processing that suggests that even though idioms are pre-stored in the lexicon, during processing, the internal words may still be activated as a type of mediation to retrieve the figurative meaning (Sprenger, Levelt, & Kempen, 2006; Titone & Connine, 1999; Abel, 2003).

On the other hand, the ratings for predictability for GYYs were moderately correlated with the ratings for literality, but not correlated at all with those for compositionality. For CYs, predictability was intercorrelated with neither compositionality nor literality. This finding replicates that observed by Titone and Connine (1994). In Li, Wang, and Zhang's (2016) study, predictability was also only moderately correlated with compositionality and literality. In all three studies, a strong correlation between predictability and familiarity was found. This result suggests that if the idioms are highly familiar to speakers, the final word of the idiom will be highly predictable regardless of the structural or semantic transparency. This finding also supports the holistic hypothesis that the most frequent idioms are more likely to be pre-listed as whole units in the mental lexicon (Swinney & Cutler, 1979; Gibbs, 1980; Cacciari & Glucksber, 1991).

The overall results regarding the relationship between the dimensions replicate those reported by Li, Wang, and Zhang(2016) in their norming study on Chinese idioms. The major difference between the two studies is the test material. The idioms in the present study are of two types, CYs and GYYs, while Li et al. only investigated CYs. In addition, the selected idioms in the study are of the top frequency band in a large-scale corpus, while the material in Li et al.'s study was selected based on syntactic structure and only 16 items occur in both of our lists. Despite the different material, the intercorrelations between the ratings on the dimensions for CYs are basically the same. This finding suggests that it is idiomaticity rather than frequency that plays a more determining role in idiom comprehension.

Syntactic structures and norms

To examine whether the different ratings on the six dimensions is influenced by syntactic structure, ratings/scores for the idioms for the six dimensions were analyzed by one-way analysis of variance with Brown-Forsythe statistics assuming unequal sample size (Tomarken & Serlin, 1986) with the ratings/scores of the six dimensions as dependent variables and the syntactic structures as the single factor. Because ratings on the majority of the dimensions were negatively skewed, the ratings were transformed by Log10 algorithm and then bootstrap transformed. Table 3.9 displays the results.

Table 3.9: Coefficients of the six dimensions of under the influence of syntactic structure

	GYY				**CY**			
	F	df_1	df_2	Sig.	F	df_1	df_2	Sig.
FAM	1.874	3	8.077	.212	1.441	6	187.779	.201
MEA	1.649	3	26.448	.202	1.735	6	176.416	.115
COM	4.037	3	29.193	.016	2.287	6	213.964	.037
LIT	3.903	3	8.170	.054	2.010	6	196.557	.066
PRE	1.170	3	7.436	.384	3.506	6	139.143	.003
FOR	6.245	3	41.116	.001	1.710	6	209.281	.120

For GYYs, there are four syntactic structures (df_1=3); for CYs, there are seven syntactic structures (df_1=6).

For GYYs, syntactic structure has significantly influenced the ratings on compositionality, literality (marginal), and linguistic register, as illustrated in Figures 3.2 and 3.3.

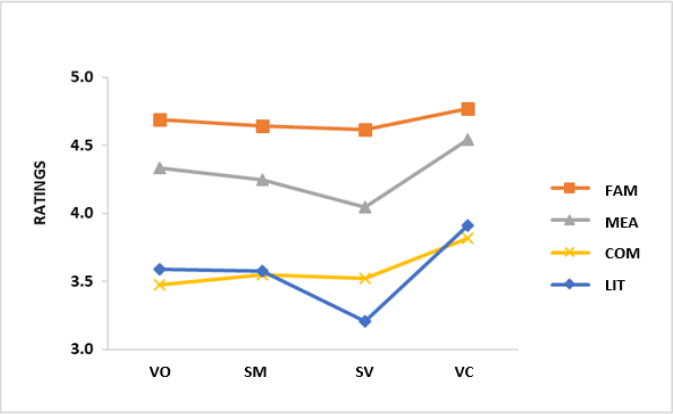

Figure 3.2: Ratings for familiarity, meaningfulness, compositionality, and literality by GYYs' syntactic structures

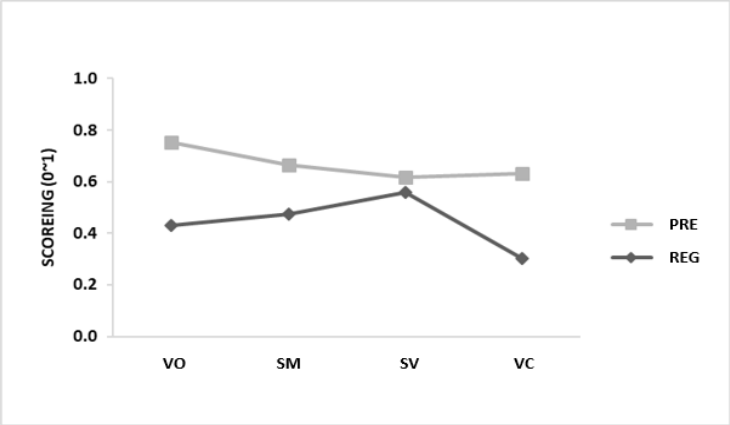

Figure 3.3: Scores for predictability and register by GYYs' syntactic structures

Pairwise comparisons showed that ratings for VO structures (e.g., 打官腔 *da-guan-qiang* play-bureaucracy-tone 'speak in a bureaucratic tone') and VC structures (e.g., 靠不住 *kao-bu-zhu* lean-not-stable 'unreliable; unstable and cannot be leaned on') were significantly different for compositionality (p=.023) and literality (p=.020), with VC being rated more highly than VO. This result is not surprising, as for many VC GYYs (e.g., 靠不住 *kao-bu-zhu*), the literal meaning and the metaphorical meaning are both frequently used, but the VO GYYs only have

metaphorical meanings. For register, a significant difference appeared between the VO structures and VC structures (p=.001) and between the SM structures and VC structures (p<.000), VC GYYs being rated the more informal within both pairs. This result suggests that native speakers use more VC idioms often in their daily communication.

For CYs, compositionality, literality (marginal), and predictability are influenced by syntactic structures, as illustrated in Figures 3.4 and 3.5.

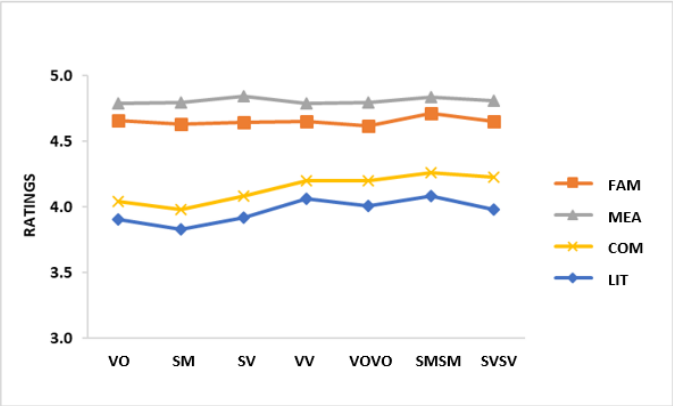

Figure 3.4: Ratings for familiarity, meaningfulness, compositionality, and literality by CYs' syntactic structures

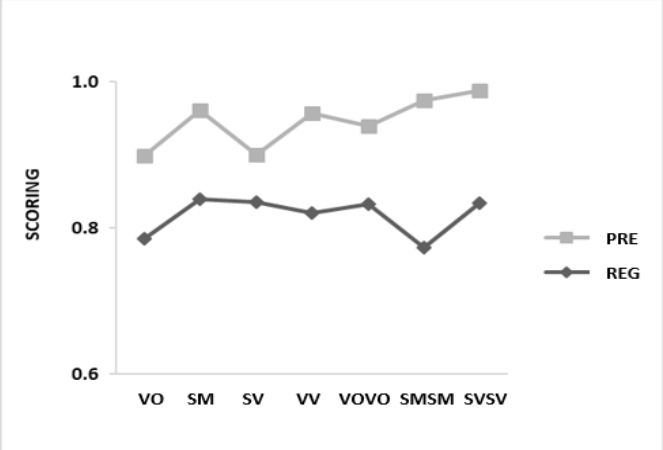

Figure 3.5: Scores for predictability and register by CYs' syntactic structures

Pairwise comparisons showed that ratings for SM structures (e.g., 花花世界*hua-hua-shi-jie* colorful-colorful-life-world 'the dazzling human world with its myriad temptations') and SMSM structures (e.g., 五花八门*wu-hua-ba-men* five-pattern-eight-category 'of a wide variety')

were different for compositionality (p=.057) and literality (p=.045), with SM CYs being rated lower than SMSM. The results for compositionality and literality replicate those observed by Li, Wang, and Zhang (2016). In their study, the ratings for literality for CYs with an SM structure were significantly lower than for those with VO, VV, VOVO, SMSM, and SVSV structures, indicating that SM CYs are more difficult to interpret literally than CYs with the other five structures. Regarding compositionality, Li et al. also reported that CYs with an SM structure were rated significantly lower than those with VO, SV, VV, and VOVO structures, which also suggested that SM CYs are less semantically transparent than CYs with other structures. With respect to predictability, the difference between ratings for SV structures (e.g., 热血沸腾 *re-xue-fei-teng* hot-blood-boiling-rise 'burning with righteous indignation') and SMSM structures is significant (p=.042), indicating that SV CYs are more predictable than SMSM CYs. For an SV CY, the four constituent words form a whole syntactic phrase, and the choice of the last word is constrained by the accumulating semantics of the previous three characters, which increases the predictability. For an SMSM CY, the four constituent words construct two phrases, technically speaking, two independent phrases in terms of syntax. The prediction of the last word is only constrained by the word immediately before it. Therefore, the word choice could be more open. However, this result was not found in Li et al.'s study. The current study also failed to replicate the significant results for familiarity and meaningfulness that Li, Wang, and Zhang (2016) have. found. We speculate that this is due the high frequency of the idioms in our study overriding the influence of the syntactic structures. Another reason for the inconsistency of the results between the two studies is that the number of the idioms was not equal for each syntactic structure because the choice of material did not allow for controlling that parameter.

Summary

This norming study provided normative results for six linguistic dimensions of measurement (familiarity, meaningfulness, literality, compositionality, predictability, and linguistic register) for 425 high-frequency idioms of two different structure types. This study made several novel contributions, including that we 1) included GYYs, the more colloquial of the two styles of idiomatic expressions under consideration, in our examination; 2) provided a comparison between two types of idioms and found different distribution patterns in terms of native speakers' intuitions on the aspects listed above; and 3) examined linguistic register dimension, which may also impact the comprehension and processing of idioms.

In summary, idioms are multifaceted expressions, as has been demonstrated by this study and others. Although the focus of this study was to investigate the factors that could potentially influence the online processing of idioms, other factors, such as emotional valence and figurative genesis, may also play a role in determining how speakers perceive, process, and use idioms. It is our hope that a more comprehensive study will include norms for these facets in order facilitate research pertaining not only to psycholinguistics but also to language learning and teaching.

CHAPTER 4: STUDY 2

LEXICAL ACTIVATION IN THE PROCESSING OF TWO TYPES OF CHINESE IDIOMS BY L1 SPEAKERS

Introduction

Whether idioms are 'words' or 'phrases' has been a central topic in the research on idiom processing and comprehension. Words and phrases are assumed to have different lexical representations in the lexicon, and the status of representation directly influences processing and comprehension (Abel, 2003). In terms of semantics, idioms have similar characteristics to words. For example, the opacity and nonliteralness of idiom meanings are analogous to the arbitrariness of word meanings, and as such it is logical to assume that an idiom form together with its figurative meaning is holistically stored in the lexicon. This view, was labeled by Swinney and Cutler (1979) as the lexical representation hypothesis, assumes that readers begin activating the whole form as a single word as soon as they encounter the first constituent word idioms. From this perspective, idioms are noncompositional, which means that the overall meaning of an idiom is not related to the meaning of constituent words. To understand an idiom, speakers must have some prior knowledge or experience with it (Bobrow & Bell, 1973; Swinney & Cutler, 1979; Gibbs, 1980; Cacciari & Glucksber, 1991). On the other hand, where the structures of idioms are concerned, as multiword configurations idioms are more akin to phrases whose structures are decomposable and whose meanings receive more or fewer contributions from their constituent words. This view, often referred to as the configuration hypothesis (Cacciari & Tabossi, 1988), sees idioms as a configuration of a string of words in the upper level of memory, and as a reader encounters an idiom, each constituent word will be lexically accessed at a lower level and feedback the activation of the whole form. However, once a phrase is identified as an idiom by

speakers who identify unique information (e.g., an idiom 'key') literal processing stops, and the activation of the literal interpretation of the idiom is inhibited (Cacciari & Tabossi, 1988). According to this perspective, idioms are not homogeneous in nature, and thus, for idioms with different degrees of decomposability, the literal meaning of the constituent words is activated to different extents (Cacciari & Tabossi, 1988; Cacciari & Glucksberg, 1991; Nunberg, Sag & Wasow, 1994; Gibbs, Nayak, Bolton, & Keppel, 1988; Glucksberg & Keysar, 1993; , Sprenger, Levelt & Kempen, 2006). Alternatively, a hybrid view put forward by Cutting and Bock (1997) argues that an idiom serves as a conventional lexical concept, is mentally represented in one layer of the lexicon, and constituent elements are activated in another layer (the lemma level) in the ways that speakers link multiword configurations to their conventional lexical concept. One thing that the hybrid view and configuration hypothesis both predict is that some degree of literal activation is achieved before the retrieval of the figurative meaning (Holsinger, 2013).

The competing theories have produced a considerable amount of empirical research, from which mounting evidence has shown that the semantic activation of constituent words does occur (e.g., Cutting & Bock, 1997; Sprenger, Levelt & Kempen, 2006). However, it remains an open question at what stage internal activation occurs and whether activation differs across different types of idioms. The present study investigates internal semantic activation to two types of Chinese idioms in early stages of processing. The study has two aims: (1) to examine whether the constituent words of idioms are semantically activated in the processing of visually presented idioms, and (2) to compare the lexical nature of the two types of idioms through evidence gathered from their processing by native speakers. The chapter is structured as follows. First, a review the empirical research on internal lexical access in idiom processing is given. Then, the two types of Chinese idioms are introduced in regard to their categorization criteria and basic

linguistic and pragmatic functions. Next, the main experiment, a primed lexical decision task, and its results are presented. The processing patterns revealed for the two types of idioms are compared and discussed. Finally, a general discussion is given on how our findings for Chinese idioms can contribute to the debate on the lexical nature and processing of idioms in general. Based on these findings, we argue that the categorization of Chinese idioms has psycholinguistic grounding that reflects the different lexical representations of GYYs and CYs.

Evidence of Lexical Activation in Idiom Processing

This section provides a selective review of previous studies that investigate the relationship between constituent words and the whole form using priming paradigms. Cutting and Bock (1997) employed speech-error elicitation methods with priming paradigms to investigate whether speakers are sensitive to internal syntactic and semantic information. In prepared tasks, paired idioms contained matched or unmatched syntactic structures or constituent words were visually presented for a short period to speakers who were later asked to produce one of the two idioms they had seen. Production latencies and blend errors were assessed. The authors found both the syntactic structures and literal meanings of internal words to be active during the production of both compositional and non-compositional idioms, suggesting that both types of idioms have the same mental representation. Additionally, based on the findings the authors argued that each idiom may have its own lexical concept representation in the lexicon and that concept representation can trigger the activation of constituent words at the same time. Competition between literal and figurative meanings may create semantic blend errors.

Inspired by the hybrid view, Sprenger, Levelt, and Kempen (2006) proposed that there is a superlemma layer that can mediate the retrieval of idioms' figurative meanings and the

activation of constituent words' literal meanings through three cue-production experiments with a priming paradigm. Native speakers of Dutch were asked to learn a list of phrases including some idiomatic expressions (e.g., . . .*viel buiten de boot* "to be excluded from something") and their matched novel phrases (. . .*ging met de boot* "took the boat") and to later recall a phrase upon listening to a word that repeated the noun included in a pair of idiomatic and literal phrases (e.g., *de boot* "boat"). The results showed that the repetition accelerated idiom production more effectively than literal phrase production, indicating that the internal word facilitates the activation of the whole idiom. In Experiment 2 speakers were asked to produce the final word of an idiomatic or literal phrase fragment primed by a word that was either semantically related, phonologically related, or unrelated to the to-be-produced final word. The authors found that an idiom can be primed by a word that is semantically related to only one of its content words. In Experiment 3, speakers were told to that they would see two types of tasks: producing the final word of an idiom fragment when a question mark appeared after the presentation of the idiom fragment; naming a given word when a word appeared after the presentation of an idiom fragment. The given word was related or unrelated to the missing word in the idiom. The results showed that preparing to produce the final word primed the naming of both phonologically and semantically related words, which indicates that literal word meanings become active during idiom production.

To test the hybrid model in the domain of idiom comprehension rather than idiom production, Holsinger (2013) conducted two eye-tracking movements. He embedded an idiom phrase (e.g., kick the bucket) and its novel counterpart (e.g., kick the pail) into contexts that are syntactically compatible (e.g., John **kicked the bucket** last Thursday) or incompatible (e.g., It was surprising to see someone as skilled as John completely miss the ball when he **kicked. The**

bucket full of orange slices was destroyed when he accidentally missed the ball) or idiomatically biased (e.g., Swimming with sharks is a dangerous and unpredictable profession. As a result of the shark attack several oceanographers **kicked the bucket** last Thursday evening) or literally biased (e.g., John spent all day filling things with cement as a nasty prank. Several people broke their toes when they **kicked the pail** last Thursday evening and may sue). Participants were not assigned any specific task, but their eye movements were tracked when they were presented with four words, one being idiomatically related to the idiom (e.g., death), one being literally related to a component word of both types of phrases (e.g., foot, related to "kick"), and two being distractor words. The results in terms of proportions of glances showed that participants had a preference for looking at the literally associated word (foot) in the early time window across all context conditions except in the case for the idiom-biased condition under which for more time in the early stage, the idiomatically associated word was looked at (death). This research provides support for the assumption that some level of literal activation of constituent words does occur in early stages of idiom processing prior to the retrieval of the figurative meaning.

Cacciari and Tabossi (1988) employed a primed lexical decision task to study the literal or idiomatic activation of the last constituent word during idiom processing. Idioms (e.g., "in seventh heaven") and matched novel phrases (e.g., "in seventh position") were embedded in neutral sentences (e.g., "After the excellent performance, the tennis player was **in seventh heaven**" versus "After the excellent performance, the tennis player was **in seventh position**"). The sentences were auditorily presented and used to prime visually presented target words that were idiomatically related to the whole idioms, literally related to the last words of phrases (idioms and non-idioms) or unrelated. Speakers were asked to make lexical decisions on the target words. The results show that with predictable idioms, subjects were faster to judge

idiomatically related targets than literally related targets; with unpredictable idioms, subjects were faster to judge the literally related targets than the idiomatically related targets. When the experiment was replicated with target words presented 300 ms after the end of an idiom was heard, subjects were equally fast at identifying idiomatically and literally related targets. These results suggest that the activation of the figurative meaning is an ongoing process and that before an idiom is identified as an idiom, the internal meanings of constituents are activated, and one determining factor of how soon an idiom can be recognized is the predictability of the idiom. These findings also lead the authors to draw an analogy between idioms and polymorphemic words whereby any configuration cannot be recognized before a sufficient amount of information has been received.

Employing the similar cross-modal priming paradigm, Titone and Connine (1994) examined the influence of the predictability and literality of idioms on contextualized comprehension. Idioms with high (e.g., 'bury the hatchet') and low (e.g., 'hit the sack') levels of predictability were embedded in auditorily presented sentences and used to prime for idiom-related words (e.g., 'bury the hatchet' – forgive) or unrelated target words (e.g., 'bury the hatchet' – gesture). Their results reveal a priming effect for both highly predictable and less predictable idioms priming for idiom-related visual targets when targets are presented in the idiom offset position (after the whole idiom is heard). However, priming was found to be more significant for highly predictable idioms than for less predictable idioms when the visual target was presented before the whole idiom was heard. The activation of the literal meaning of the idiom-final word was found for all types of idioms except in the case of highly predictable-nonliteral idioms. This finding seems to suggest that literal activation is likely to occur throughout the comprehension process rather than only in early stages of processing. Also with

the focus on discussing the idiom processing at different stages, in 2014, the two authors employed a cross-modal primed lexical decision task to investigate whether the activation of figurative versus literal meaning of an idiom is modulated by the idiom's decomposability, familiarity, and literal plausibility. In the experiments, participants listened to an idiom-carrier sentence and then made lexical decisions on a word that is related to the literal meaning of the final word in the idiom, or the figurative meaning of the whole idiom, or unrelated to the idiom in any way. The results showed that figurative meaning activation steadily increased as the idiom string unfolds until 1000 ms later of the auditory presentation. Different linguistic factors of idioms affected the activation of figurative interpretation also in a time-dependent fashion. Across the two experiments, the author concluded that multiple linguistic factors constrain idiom comprehension, and the magnitude of figurative activation varies on different processing stages. The focus of the present study is also the early stage of idiom processing. According to the previous findings, we anticipate to find the literal activation of constituent words in native speakers' processing of both types of Chinese idioms.

Two Types of Chinese Idioms

The two types of idioms this study investigates are three-word[8] idioms often referred to as *guan-yong-yu* (GYY) 'conventional-use-language' (e.g., 飞毛腿 fly-feather-leg "fleet-footed runner") and four-word idioms otherwise referred to as *cheng-yu* (CY) "fixed-language" (e.g., 画蛇添足 draw-snake-add-foot "to ruin the effect by adding something superfluous"). Like all idiomatic expressions, they both have figurative meanings that are not always derivable from the

[8] For the purpose of distinguishing internal elements and the whole idiom form and to compare our results with previous findings, in this study the constituent elements of Chinese idioms are considered as words rather than characters or syllables.

meanings of their constituent words. However, in the considerable body of literature on Chinese idiom processing, much research has focused on CYs (Zhou, Zhou & Chen, 2004; Liu, Li, & Shu et al., 2010; Chung, Code & Ball, 2004; Zhang, Yang & Gu et al., 2012; Cacciari, Padovani & Corradini, 2007) which little attention has been paid to GYYs. Admittedly, CYs are ideal targets for research focusing on the lexical nature of idioms because CYs have a homogenous format. They are all composed of four characters and allow for no semantic substitution or syntactic variations. In addition, CYs quantitatively prevail in Chinese with roughly 97% of idioms conforming to the four-character format (Liu, Li, Shu, Zhang & Chen, 2010). However, the primary reason that CYs and GYYs are distinguished in the Chinese tradition pertains to their figurative genesis and linguist register. CYs originated from the classical Chinese language with traceable sources. According to Xiao's (1987) statistics, more than 60% of CYs can be dated to the pre-Qin Dynasty (~ 221 A.D.). Therefore, CYs carry a considerable amount of classical vocabulary and grammar not compatible with contemporary Chinese. Due to their origins, CYs are considered a high-end formal language and are often used in written discourse. GYYs, on the other hand, are folk creations stemming from people's life and working experience and often carry a casual and sarcastic tone (Wen, 1989). Most GYYs also have a flexible structure, allowing for lexical addition and substitution as well as syntactic operations, such as those of aspectual affix insertion or dislocation. GYYs are considered colloquial language and are often used in spoken and informal contexts (Wen, 2007).

To our knowledge, no research has compared the two types idioms from a cognitive perspective. In Study 1, we conducted a large-scale norming study collecting native speakers' ratings on the 425 most commonly used GYYs and CYs on five linguistic dimensions that have been claimed to be able to influence idiom processing in the literature: familiarity (how often an

idiom is encountered), meaningfulness (how well a speaker understands the figurative meaning of an idiom), compositionality (to what extent constituent words are semantically related to the figurative meaning of an idiom), literality (the likelihood of the literal interpretation of an idiom being realized in the real world), and final word prediction (the proportion of an idiom fragment that is completed idiomatically). The results show that CYs are scored higher than GYYs on every dimension with predictability and compositionality being the most different dimensions. In this study we attempt to investigate whether the two types of idioms differ from one another in their lexical representations.

The Present Study

Research question and design

The present study sets out to investigate whether the internal words of an idiom are semantically activated in early stages of the visual processing of Chinese idioms using a primed lexical decision task. The semantic relation between prime and target words was manipulated to examine (a) whether priming effects would be observed between an idiom prime and a target word semantically related to a specific constituent word in this idiom and (b) if priming effects (or lack thereof) would be observed for both GYYs and CYs. In addition, the second constituent word of an idiom was chosen as the word of interest because (a) the second position is a non-boundary position for both GYYs and CYs, and so it serves the purpose of examining internal lexical access while (b) the second word is presented earlier than the third word in CYs (four-word idioms), thus rendering it appropriate for investigating early stages of idiom processing. The prime involves three conditions: (a) the related idiom condition with the prime being an idiom (e.g., 风花雪月 wind-flower-snow-moon ‘romantic theme’) whose second constituent

word (e.g., 花 ‘flower’) is semantically related to the target word (e.g., 園林 park-woods ‘garden park’), (b) the related novel condition with the prime being a rule-generated novel phrase (e.g., 卖花姑娘 sell-flower-aunt-mother ‘flower girl’) containing the same second constituent word (花 ‘flower’) related to the same target word (園林 park-woods ‘garden park’), and (c) the unrelated idiom condition with the prime being an idiom (e.g., 勇往直前 courage-toward-straight-forward ‘march forward courageously’) that does not contain a second constituent word semantically related to the same target (園林 park-woods ‘garden park’). Table 4.1 shows the design and example stimuli for both types of idioms.

Table 4.1: Design and example stimuli with related constituent words underlined

GYY		
Prime Condition	Prime Stimulus	Target DSW
Related idiom	敲竹杠 knock-bamboo-lever “take advantage of someone’s being in a weak position”	熊貓 bear-cat ‘panda’
Related novel	撑竹竿 prop-bamboo-pole “prop up a bamboo pole”	
Unrelated idiom	做人情 make-people-affection “give someone a favor”	
CY		
Prime Condition		Target DSW
Related idiom	风花雪月 wind-flower-snow-moon “romantic themes”	園林 park-woods ‘garden park’
Related novel	卖花姑娘 sell-flower-girl-mother “flower girl”	
Unrelated idiom	勇往直前 courage-toward-straight-forward “march forward courageously”	

Dependent variables include the amount of time participants took to make lexical decisions and the response accuracy of the same set of target words observed under the three prime conditions. Predictions made are outlined in the following section.

Predictions

Regarding semantic priming effects observed across the three conditions for each type of idiom, three patterns are predicted.

Prediction #1: If the internal words of idioms are not semantically activated during processing, no priming effect should be found for the related idiom condition, and then we predict that participants' responses to the target words under the related idiom condition should not be significantly different from responses made under the unrelated idiom condition but would be significantly poorer than those of the related novel phrase condition. The priming effect pattern observed under the three conditions should thus be as follows: Related novel > Related idioms = Unrelated idioms.

Prediction #2: If the internal words are activated during idiom processing, then significant priming effects should be found for the related idiom condition rather than for the unrelated idiom condition. If the magnitude of internal activation is correlated with the compositionality of the phrase containing the word of interest, related novel phrase primes whose compositionality is higher should have a strong facilitating effect on targets than the related idioms with less compositionality. Then we predict that the following priming effect pattern should emerge: Related novel > Related idioms > Unrelated idioms.

Prediction #3: If the related internal words are semantically activated regardless of the compositionality of phrases, responses made under the related idiom condition should not be significantly different from those made under the related novel phrase condition, and both conditions should show significantly more priming effects than the unrelated idiom condition. Priming effects observed under the three conditions should thus follow a following pattern: Related novel = Related idioms > Unrelated idioms.

To compare the three prime conditions, lexical properties of the three types of primes must be controlled. To observe semantic priming effects, the semantic relationship between the prime and target must also be controlled. For these reasons, two norming studies were conducted and are presented in the next section.

Materials

The selection of primes

The lexical properties of three prime phrases are controlled to ensure that if a difference is found between the three conditions, this should be the result of participants utilizing different processing mechanisms upon reading different types of primes but not of differences in other lexical dimensions between the three primes.

Thirty GYYs and thirty CYs were first selected from Study 1's database based on the following criteria: (a) average ratings on familiarity and meaningfulness for the select idioms must above point 4; (b) the second constituent word of each idiom must be nominal in contemporary Chinese (numerals excluded); (c) an idiom whose second constituent word occurs twice in this idiom (e.g., 自欺欺人 self-deceive-deceive-people "to deceive oneself as well as other") is not included; (d) the thirty idioms of each type must have thirty distinctive second constituent words.

Second, another thirty GYYs and CYs were selected from Study 1 as unrelated idioms. The average ratings on familiarity and meaningfulness must also be greater than 4, and stroke numbers are matched for each pair of related and unrelated idioms. Table 4.2 shows descriptive statistics and paired-sample t-test results for the familiarity ratings, meaningfulness ratings, and stroke numbers.

Table 4.2: Descriptive statistics and one-way ANOVA results for related and unrelated idioms

		Familiarity	Meaningfulness	Stroke numbers
GYY	Mean of related idioms	4.44	4.82	23.27
	Mean of unrelated idioms	4.49	4.80	23.30
	F-value	.26	.97	.00
	p-value	.61	.33	.98
CY	Mean of related idioms	4.61	4.79	26.80
	Mean of unrelated idioms	4.66	4.81	26.90
	F-value	1.13	.43	.00
	p-value	.29	.52	.95

Finally, we selected related novel phrases from a Chinese written text corpus BCC (BLCU Corpus Center; Xun, Rao, Xiao, & Zang, 2016). We did not use homemade novel phrases to ensure the legitimacy of the novel phrases and to ensure that they were actually used in the real world. The BCC corpus offers a 'fuzzy search' function with which we searched for a four-character sequence with the second word specified and with words in the other positions left unspecified (e.g., entering a sequence with format ".手.."). The procedure yielded all sentences containing the ".手.." sequence as well as statistics on the token frequency of each ".手.." sequence. Based on these statistics, we selected top-ranked novel phrases having similar configurations with their related idiom counterparts as potential candidates. Because all related idioms were selected from the top frequency bin and were rated as very familiar, even though the novel phrase candidates are highly ranked in the corpus, it is difficult to match frequencies for a novel phrase and an idiom counterpart. Because the ultimate purpose of controlling the frequency and stroke numbers of the two types of primes is to ensure that one type is not considerably more difficult to recognize than the other type, we adopted the approach used by Tabossi, Fanari, and Wolf (2008) for their third experiment, which involved a series of phrase recognition tasks used to determine if the RTs of the chosen related idioms would match the RTs of their chosen novel phrase counterparts. Twenty Chinese college students who had not participated in the main experiment were recruited over two rounds of recognition tasks (10

people in each round). They were asked to judge whether the phrase they saw on the computer screen was acceptable Chinese language by pressing the corresponding keys (A for "Yes" and L for "No"). Participants were asked to respond as rapidly as they could. The material was split into two counterbalanced lists, which means that the same participant would only see one item of each idiom-novel phrase pair. The RT data were log transformed and analyzed with one-way ANOVAs. The results showed no significant difference between idioms and novel phrases for RT data of the CY group, $Mean_{-idiom}$=2.824, $Mean_{-novel}$ =2.854, $F(1, 149)=2.710$, p=.102, or for the judgment accuracy of the CY group, $F(1, 149)= 0.336$, p=.563. For the GYY group, RTs also did not return a significant difference between the idioms and their matched novel phrases, $Mean_{-idiom}$=2.850, $Mean_{-novel}$ =2.845, $F(1, 149)=.285$, p=.594, or for judgment accuracy, $F(1, 149)=1$, p=.319.

In summary, the results of the norming study show that the selected novel phrases should not be more difficult to recognize than the selected idioms. The procedure confirms that all of the selected novel phrases are acceptable phrases in Chinese.

The selection of target words

All target words are disyllabic compound words and are chosen based on the following criteria. First, a target DSW is semantically related to the second constituent words of the related condition primes (related novel phrases and related idioms share the same second constituent words). Second, a DSW cannot contain or be semantically related to any other constituent words of its related primes. Third, a DSW cannot contain or be semantically relate to any constituent words of its unrelated prime. Fourth, efforts were made to avoid using DSWs whose characters share radicals with related constituent words in primes.

To ensure the presence of a semantic relationship between the second constituent words and the target DSWs, a norming study was conducted to examine if priming effects could be found when the chosen target DSWs were primed by the related second constituent words presented alone. In the norming study, the second constituent words of related primes (e.g., 花 "flower" in 风花雪月 wind-flower-snow-moon 'romantic themes') and those of unrelated primes (e.g., 往 'toward' in 勇往直前 courage-toward-straight-forward "march forward courageously") were isolated and paired with a DSW (e.g., 園林 park-woods "garden park") to form a related prime-target pair (e.g., 花-園林 flower-garden park) and an unrelated prime-target pair (e.g., 往-園林 toward-garden park). A paired-presentation LDT was adopted (McRae & Boisvert, 1998). For each trial, a monosyllabic prime word (in simplified characters) and a target DSW (in traditional characters) were presented side by side on the same screen. The participants' task was to judge if both words are correct in Chinese by pressing corresponding keys (A to denote that they are both correct and L to denote that they are not both correct). Figure 4.1 presents a typical trial of this norming priming experiment.

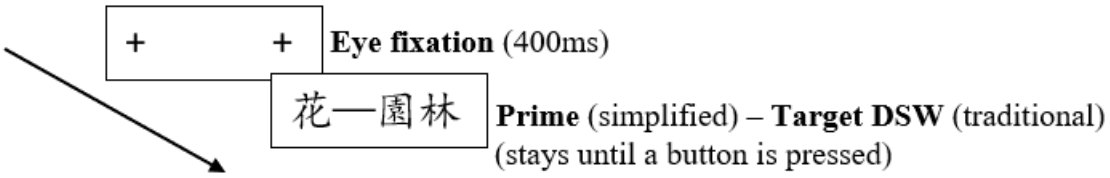

Figure 4.1: Paired presentation of a trial of the norming priming experiment

As can be seen in Figure 4.1, in a paired presentation, each trial consists of two screens. The first screen presents the fixations, and the second screen presents both the prime and target until a button is pressed. Paired-presentation methods were found to have backward priming effects. Backward priming effects indicate that subjects evaluate the association between the prime and target after the target's presentation (Neely & Keefe, 1989; Sbelton & Martin, 1992).

We intended to utilize backward priming effects to inhibit the activation of a competing meaning that a prime may carry. In Chinese, some characters carry more than one meaning, and all of these meanings can be used frequently and compete with each in the word recognition (Gaskell & Marslen-Wilson, 2002; Rodd, Gaskell & Marslen-Wilson, 2004). For example, the character 花 can serve as a noun meaning 'flower' or as a verb meaning 'to spend,' both of which are frequently used in communication. When the character is present in a given context such as in an idiom (e.g., 风花雪月 wind-flower-snow-moon 'romantic themes'), the character 花 can only denote one meaning, 'flower,' which is related to the target DSW 園林 'garden park'. We argue that the target word of a paired-presentation can also serve as such in relation to the prime character. When the prime-target pair is perceived as semantically related, the priming effect should be enhanced, and if the prime-target pair is perceived as unrelated, the priming effect should be inhibited, and thus we can draw a firm conclusion on whether target words are truly related to words of interest in primes. Another twelve students who neither participated in the phrase recognition task or in the main priming experiment were recruited in the current norming study. They were given a counterbalanced list of trials whereby one participant was not shown the related prime-target pair or its unrelated counterpart. RT data were Log transformed with error data removed. Significant priming effects were found for the related pairs for GYYs, $F(1, 152)=20.72$, $p<.000$, with the related target words being recognized 211 ms faster on average than the unrelated target words, and CYs, $F(1, 175)= 14.67$, $p<.000$, with the related target words being responded to 115 ms faster on average than the unrelated target words. In terms of judgment accuracy, GYY data reveal the marginal significance of the related pairs, $F(1, 179)= 2.94$, $p=.088$, and CY data show a significant priming effect for the related pairs, $F(1, 179)=5.11$, $p=.025$. This result confirms the semantic relationship between the selected target DSWs and

words of interests included in the related primes. Appendix D presents a full list of the test material.

Fillers and counterbalancing

Finally, a set of 120 filler trials were made up, including 30 correct prime-target filler trials with prime and target both being correct and 90 incorrect prime-target trials with prime and target both being incorrect. The experimental test items were divided into three counterbalanced lists such that one participant would see the same target DSW primed by two different conditions. For example, participants presented with a related-idiom-target pair 风花雪月-園林 (wind-flower-snow-moon "romantic themes" – park-woods "garden park") could not see the related-novel-target counterpart 卖花姑娘-園林 (sell-flower-girl-mother "flower girl" – park-woods "garden park") or the unrelated-idiom-target counterpart 勇往直前-園林 (courage-toward-straight-forward "march forward courageously" – park-woods "garden park"). The 120 fillers were repeatedly used in three counterbalanced lists. The total number of trials for each list was set to 180 (30 GYY trials + 30 CY trials + 30 correct filler trials + 90 incorrect filler trials). Each list was split into two blocks with one block including 30 GYY trails and 60 filler trials and with the other including 30 CY trials plus another 60 filler trials. The two blocks were separated by two-minute intervals. In the experiment, each participant was pseudo-randomly assigned to one of the three lists.

Procedure

The study employed a priming paradigm to a lexical decision task. This research technique is designed to investigate the morphological, phonological, semantic, or orthographic associations between the prime and target words/phrases (Evett & Humphreys, 1981; Forster & Davis, 1984). Under the priming paradigm, participants are presented with a prime word/phrase,

which is displayed for a controlled amount of time of (e.g., 50 ms to 2000 ms) before presenting a target word/phrase on which lexical decisions are made. The prime is sometimes preceded by a sequence of masks to prevent participants from be aware of the presentation of primes to investigate early word processing. However, masks are not used in the present study. The aim of the study is to test whether a constituent word of an idiom is semantically activated during idiom processing and in turn to provide semantic cues on a related target word. Therefore, participants are supposed to see prime idioms or control phrases consciously. In the following sections we first present the experimental procedure and then illustrate how we determined certain important parameters of the procedure.

In the experiment, each participant was tested individually in a quiet computer room and was seated in a chair positioned approximately 60 centimeters from a computer screen. The presentation of stimuli and recording of responses were programmed by Paradigm 2.4 (Tagliaferri, 2008). For each trial, a fixation '+' was presented for 300 ms on the center of the screen. A prime sequence was then presented in simplified Chinese characters (in 16-point Kati font) in the center of the screen for 400 ms. Next, the prime sequence was replaced with a target disyllabic word (DSW) in traditional Chinese characters (in 16-point Kati font). Subjects were instructed to press one of two buttons on the keyboard (A for "Yes" and L for "No") to indicate whether the words shown in traditional characters were Chinese words. Participants were also asked to make a decision as quickly and accurately as possible. After a key was pressed, a message stating "press spacebar to continue" appeared to mark the start of the next trial. Six recall questions asking if a certain phrase had appeared in the previous trial were randomly added to the experiment. A 10-trial training session with two recall questions was conducted

prior to the main test. The whole session lasted approximately 25 minutes. Figure 4.2 illustrates how a typical trial of the main experiment proceeded.

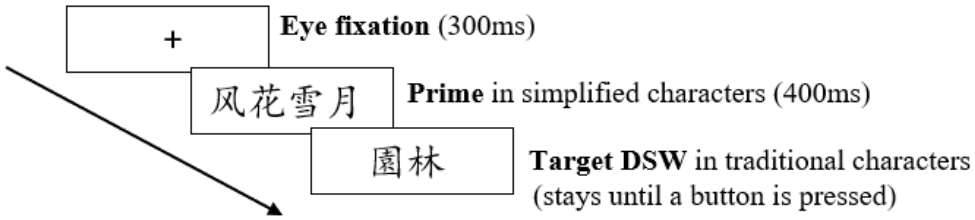

Figure 4.2: Presentation of a trial of the main experiment

Some important parameters set for the procedure were determined in the pilot phase and by referring previous research findings. First, the 400-ms stimulus onset asynchrony (SOA; the amount of time between the onset of the prime and the target) was determined in the pilot phase of the experiment. Previous research (Neely, 1976, 1977; McRae & Boisvert, 1998) has claimed that a long SOA of 400 ms or more can cause strategic processing or expectancy generation effects. The presence of such effects mean that participants notice the association between some prime-target pairs then strategically and prospectively predict the targets, resulting in amplified priming effects. However, this claim was argued to be inconclusive by Chiarello, Burgess, Richards, and Pollock (1990) who found no semantic priming when applying a 575 ms SOA in a naming task. Moreover, this claim is generally based on the results of word-to-word prime-target experiment while in the present study we used a phrase-to-word priming experiment. In the pilot phase, three SOAs (250 ms, 400 ms, and 600 ms) were attempted. Subjects reported that an SOA of 250 ms was too brief to capture a whole phrase and especially in the case of novel phrases. In Cutting and Bock's (1997) primed idiom recall study, the presentation of the prime idiom was determined by adding 40 ms for each letter in a content word and 20 ms for a functional word to a base SOA of 250 ms in order to allow participants to read each idiom only once. If using

approach and adding 40 ms for each content word and 30 ms for a function word to the base SOA 250 ms, the SOA would vary from 360 to 410 ms. Given the novel phrase condition which may take a bit long to read (although the norming study suggested no significant difference in RT for idiom and match novel conditions, novel phrases did take longer processing time), an even SOA of 400 ms was adopted all prime conditions. This medium SOA would allow time for participants to see the prime phrases while not allowing too much time for the participants to develop a strategy. Finally, to further minimize the potential for strategic processing, we added 30 correct filler trials and 90 incorrect filler trials as mentioned above intending to make predicting the purpose of the experiment difficult.

Switching from simplified to traditional characters from the prime to the target has a two-fold purpose. The first is to increase the orthographic complexity of targets and thus to limit the likelihood of ceiling effects in the LDT given that the target words are simple and common words. The second is to avoid orthographic priming effects. For a few items of the experiment, characters included in prime phrases (e.g., the character 钢 "steel" in the idiom 走钢丝 walk-steel-thread "wire-walking; fence oneself amid opposing parties") could contain the same semantic radicals (e.g., 钅 the radical denoting "metals") as characters included in the target words (e.g., the character 铁 "iron" in the target DSW 铁匠 iron-craftsman 'blacksmith'). Chinese characters of the same semantic domain are likely to share the same semantic radicals. The metal domain serves as a prototypical example – characters for all metallic chemical elements with the exception of 汞 "mercury" contain a metal radical 钅. In this case, it is inevitable that the target DSW and its related idiom prime have an orthographic overlap. This overlap could provide orthographic priming effects to the target word, which will invalidate the priming effects if so observed. However, when using traditional characters, the semantic radicals

change in physical appearance (e.g., from 钅 to 釒). and the potential orthographic repetition is reduced.

Recall questions were applied because in the pilot study we found that the participants tended to completely ignore prime phrases and to merely wait to be presented with target DSWs since their sole task was to recognize and make decisions on the target words. To verify that participants did pay attention to prime phrases, six immediate recall questions asking if a certain phrase had been seen were randomly added (cf. Schweigert, 1986). Subjects were told that the task was also to measure native speakers' short-term memory span through lexical recognition. Given the brief SOA of 400 ms, subjects were not expected to clearly remember all the prime phrases. Thus, subjects who successfully answered 50% of the recall questions were included in the subsequent analyses.

Participants

Sixty-four students who participated in neither of the two norming studies mentioned above were recruited from four Chinese universities. All of the participants are native speakers (40 female and 24 male; $Mean_{-age}$ = 22.35) majoring in a non-Chinese language or non-Chinese-literature-related subjects. In the recruitment phase, the participants were informed that they needed to be able to recognize some commonly used traditional characters. Each participant received 5 US dollars as compensation.

Results

All 64 participants have reached the 50% threshold in the recall task with the average score being 63.3%. The RT data were screened over the following steps. First, participants

whose judgment error rates were higher than 18%[9] were excluded from the analyses. Four participants' data were removed through this procedure. Second, items provided with incorrect judgments were excluded from the analysis of corresponding participants. From this approach we removed 5.2% of the GYYs' RT data and 6.3% of the CYs' RT data. Third, data points which fell three standard deviations above individual subjects' mean RT on target trials were removed; data points which were below 300 ms or above 4000 ms were considered outliers and thus excluded from the analyses. This procedure resulted in a loss of another 2.8% of GYY's RTs and of 3.3% of CY's RTs. The RT data for both types of idioms were severely positively skewed. To normalize the distribution, RT data were inversely transformed with the formula 1000/RT following Li, Jiang, and Gor (2017). Figure 4.3 shows the distribution of the transformed RT data for GYYs and CYs.

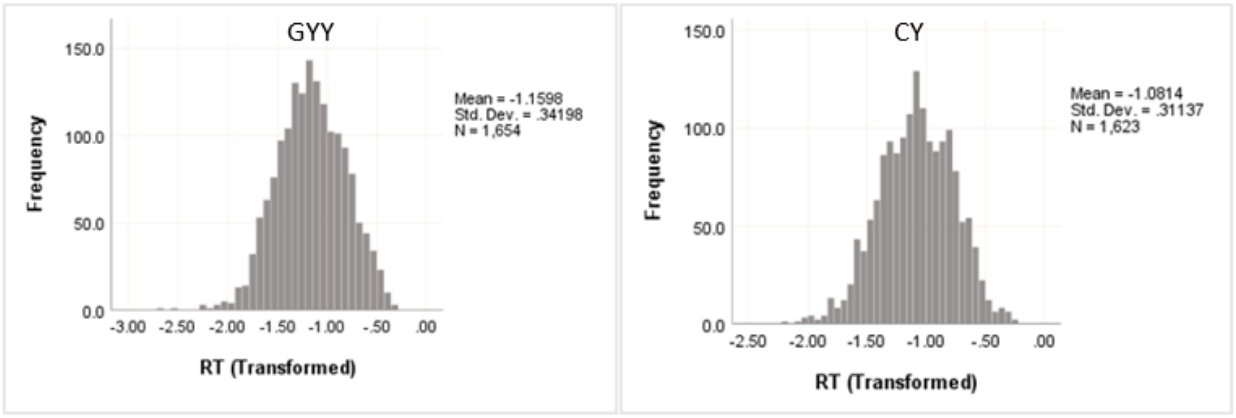

Figure 4.3: Histograms for the transformed RT data by idiom type

The RTs of each experiment were analyzed using generalized linear mixed-effects models in SPSS 25 (IBM Corp., 2017) with RTs set as the target, prime type set as the fixed factor, and subject, item, and their intercept set as random factors. The default α level was set to

[9] The 18% error rate was set through empirical-judgmental procedure, where the experimenter and a research assistant first reviewed the data distribution and made a panel decision (Berk, 1986). The 18% error rate cutoff (more than 32 items were judged incorrectly) can discriminate good performance and poor performance. The elimination will not reduce the sample size significantly.

0.05 and the confidence interval (CI) was set to 95%. Before comparing priming patterns of the two types of idioms, we first report RT results for both types and then accuracy levels for both types. Descriptive statistics of raw RTs (after employing data exclusion procedures) for each condition and for both types of idioms are presented in Table 4.3. The transformed RT data yielded a significant main effect for the Prime Type for GYYs, F (2, 1651)= 8.874, p< .000 and for CYs, F (2, 1620)= 5.224, p< .000. However, pairwise comparisons reveal different patterns for the two types of idioms.

Table 4.3: Mean RT (ms), SD (ms), and number of observations per condition by idiom type

	GYY			CY		
Condition	Mean	SD	Observations	Mean	SD	Observations
Related-novel	933	347	556	997	363	545
Related-idiom	943	327	566	1034	376	542
Unrelated-idiom	1000	406	532	1035	401	536
Total	958	362	1654	1022	380	1623

For GYYs a significant difference was found between the related-idiom condition and the unrelated-idiom condition (contrast estimate = -.053, SE=.016, t=-3.294, p=.001, CI= -.084 ~ -.021) with target words being responded to significantly faster when primed by the related idioms than by the unrelated idioms. The difference between the related-novel condition and the unrelated-idiom condition was also found to be significant (contrast estimate= -.064, SE=.016, t=-3.946, p< .000, CI= -.095 ~ -.032) with target words being responded to faster when primed by related novel phrases than by unrelated idioms. However, the significance was not reached between the related-novel and related-idiom conditions is not significant (contrast estimate= -.011, SE=.016, t=-.674, p=.5, CI= -.042 ~ .02). This pattern of priming effects corresponds with Prediction #3: the priming effect adheres to the following sequence: related novel phrases = related idioms > unrelated idioms. Figure 4.4 presents mean differences between three prime conditions for GYYs.

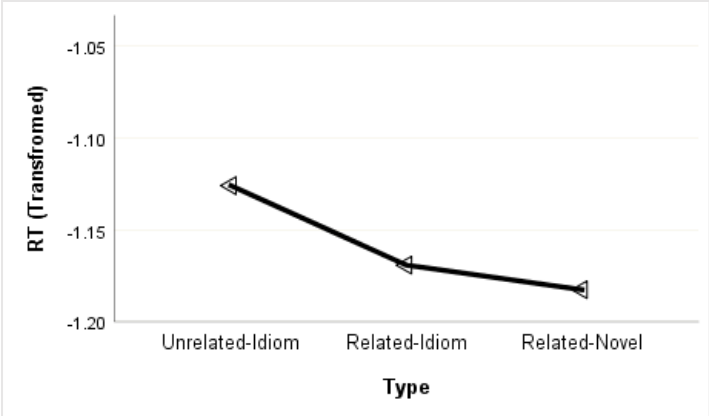

Figure 4.4: Mean of transformed RTs by prime type for GYYs

In comparison, for CYs the difference between the related-idiom condition and the unrelated-idiom condition was not significant (contrast estimate= -.005, SE=.015, *t*=-.364, *p*=.716, CI= -.034 ~ .023). However, significance was obtained for the contrast between the related-novel condition and the unrelated-idiom condition (contrast estimate= -.043, SE=.015, *t*=-2.958, *p*= .003, CI= -.071 ~ -.014) with related novel phrases showing more priming effects than the unrelated idioms. A significant difference was also obtained for the contrast between the related-novel condition and the related-idiom condition (contrast estimate= -.038, SE=.014, *t*=-2.601, *p*=.009, CI= -.066 ~ .009) with related novel phrases showing more priming effects than the related idioms. This pattern of priming effects conforms to Prediction #1, stating that the priming effect adheres to the following ranking of: related novel phrases > related idioms = unrelated idioms. Figure 4.5 presents mean differences found between three prime conditions for CYs.

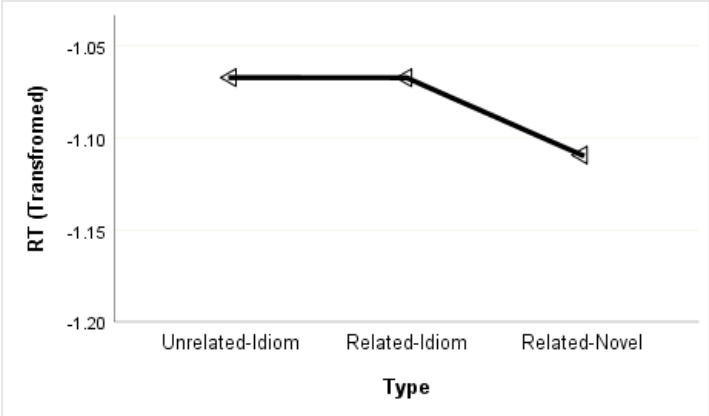

Figure 4.5: Mean of transformed RTs by prime type for CYs

Both GYYs and CYs showed main effect of prime type suggesting that the magnitude of the internal lexical activation differs across three types of primes for both types of idioms. The different patterns of priming effects found for GYYs and CYs suggest that GYYs and CYs may be processed differently. For GYYs, the related idioms patterned with the related novel phrases, revealing more priming effects on the target words than on the unrelated idioms. Because a related GYY (e.g., 敲竹杠 knock-bamboo-lever 'take advantage of someone's being in a weak position') and its novel phrase counterpart (e.g., 撑竹竿 prop-bamboo-pole 'prop up a bamboo pole') include an identical second constituent word (e.g., 竹 'bamboo') that is semantically related to the target word (e.g., 熊貓 'panda') while their unrelated GYY counterpart (e.g., 做人情 make-people-affection 'give someone a favor') does not, the presence of this priming effect indicates that the second constituent word facilitates the recognition of the target word, which further suggests that the meaning of this internal word's literal meaning is activated when participants are presented with a related GYY, the same as when they are presented with a related novel phrase. The opposite pattern was observed for CYs where related idioms patterned with unrelated idioms, showing no facilitating effect on the target words (e.g., 園林 park-woods 'garden park'), which was found for the related novel phrases (e.g.,卖花姑娘 sell-flower-girl-

mother ‘flower girl’). Since the unrelated CYs (e.g., 勇往直前 courage-toward-straight-forward ‘march forward courageously’) do not contain words (e.g., 花 ‘flower’) semantically related to the targets (e.g., 園林 park-woods ‘garden park’), it is natural that no priming effect was observed between the unrelated idioms and targets. However, by patterning with the unrelated CYs, the related CYs (e.g., 风花雪月 wind-flower-snow-moon ‘romantic themes’), while containing words semantically related to the targets, did not show priming effects on the target words. This result suggests that the literal meaning of the second constituent word of a related CY is not activated when participants are presented with the whole CY.

Regarding participants’ judgments, no data were eliminated. Table 4.4 shows judgment error rates by prime type for the GYYs and CYs. Because the judgment data are binary responses (correct vs. incorrect), binary logistic regressions were implemented through SPSS 25 with judgments set as the dependent variable, prime type set as the factor, and subject and item set as random factors. The default α level was set at 0.05, and the level of the Wald confidence interval (Wald CI) was set as 95%. The overall models were found to be significant for both GYYs (Wald χ^2 = 703, *df*=1, *p*<.000) and CYs (Wald χ^2 = 775, *df*=1, *p*<.000).

Table 4.4: Judgment accuracy, SD, and number of observations by condition for GYYs and CYs

	GYY			CY		
Condition	Mean	SD	Observations	Mean	SD	Observations
Related-novel	.965	.184	600	.930	.255	600
Related-idiom	.955	.207	600	.942	.235	600
Unrelated-idiom	.923	.266	600	.933	.250	600

Participants’ judgments returned the main effect of prime type for GYYs (Wald χ^2 = 11.045, *df*=2, *p*=.004). Pairwise comparisons of the three conditions of GYYs further show that the difference between the related-novel and the related-idiom conditions was not significant (mean difference= -.01, SE=.011, *df*=1, *p*=.377, Wald CI=-.03~.01). A significant difference was

found for the contrast between the related-novel condition and the unrelated-idiom condition (mean difference= -.04, SE=.013, *df*=1, *p*=.002, Wald CI=-.07~-.02) with the target word being judged more accurately when primed by a related novel phrase than by an unrelated idiom. Significance was also obtained for the difference between the related-idiom condition and the unrelated-idiom condition (mean difference= -.03, SE=.014, *df*=1, *p*=.021, Wald CI=-.06~.00) with target words being judged more accurately under the related-idiom condition than under the unrelated-idiom condition. This pattern is consistent with that observed from the RT data with priming effects ranked as follows: related-novel = related-idiom > unrelated idiom.

However, the main effect of prime type was not observed for CYs (Wald χ^2 =.771, *df*=2, *p*=.701). Pairwise comparisons also failed to return significance for the related-novel and related-idiom contrast (Mean Difference= .01, SE=.014, *df*=1, *p*=.409, Wald CI=-.02~.04), or for the related-novel and the unrelated-idiom contrast (mean difference= .00, SE=.015, *df*=1, *p*=.819, Wald CI=-.03~.03), or for the related-idiom and the unrelated-idiom contrast (mean difference= -.01, SE=.014, *df*=1, *p*=.551, Wald CI=-.04~.02). This pattern (related-novel = related-idiom = unrelated-idiom) is not completely the same as that revealed by the RT data where the related novel condition is significantly different from the related-idiom and the unrelated-idiom conditions, but no difference was found between the related-idiom and the unrelated-idiom conditions. Nevertheless, the judgment data show no differences between the three conditions, indicating that neither of the primes facilitated judgements. The different findings yielded by the RT and judgment data are not surprising or new. In Swinney and Cutler's (1979) study on the idiom acceptability judgment task, RT data returned significant difference between idioms and their matched novel phrases, but error rates were nearly identical for both the idioms and control novel phrases. In Li, Jiang, and Gor's (2017) study in which native speakers responded to

singular words primed by different types of compound words, the authors found a significant main effect of prime types for RT data but not for judgment data. These findings suggest that the measure of judgment accuracy may not sensitive enough to reflect the categorical difference observed when the targets that judgments are based on are too simple or familiar to native speakers.

As another possible reason that may be responsible for the insignificance of judgment accuracy, some traditional characters included in target DSWs may be too difficult to recognize, resulting in a floor effect. Therefore, following Swinney and Cutler's (1979) approach, a post hoc analysis was conducted with experimental items with an error rate of 18%[10] or higher were excluded (marked by a '*' in the list given in Appendix D). This procedure removed three GYY items and five CY items (including the related novel, related idiom, and unrelated idiom pairs). The remaining RT and judgment data were then applied to the same statistic models (generalized linear mixed-effects models for transformed RT data and binary logistic regression models for judgment data) for a second round of analyses. The results for GYYs and CYs show the same patterns as those observed before the high-error-rate items were removed.

Regarding the RT data, the main effect of prime types was observed for both CYs, $F(2, 1412)=4.202$, $p=.015$ and GYYs, $F(2, 1532)=7.357$, $p=.001$. Pairwise comparisons reveal the same patterns across the three conditions for GYYs (relate novel = related idiom > unrelated idiom) and CYs (related novel > related idiom = unrelated idiom). For GYYs, the related-novel condition patterned with the related-idiom condition ($t=-.767$, $p=.443$) and the unrelated-idiom condition was outperformed by both the related-novel condition ($t=-3.642$, $p<.000$) and the

[10] Li, Jiang, and Gor (2017) used 20% as the error rate cutoff. In this study, the cutoff was determined through empirical-judgmental procedure (Berk, 1986). The 18% error rate (more than 16 participants judged an item incorrectly) can discriminate the difficult-to-recognize target words and easy-to-recognize target words.

related-idiom condition (t=-2.894, p<.004). For CYS, the related-novel condition outperformed the related-idiom (t=-2.178, p=.03) and unrelated-idiom conditions (t=-2.745, p=.006) while the related-idiom and unrelated-idiom conditions patterned together (t=-.566, p=.572).

In regard to judgment accuracy, priming effect patterns observed for the item-deleted data were also found to be consistent with those obtained prior to item deletion for both GYYs (relate-novel = related-idiom > unrelated-idiom) and CYs (relate-novel = related-idiom = unrelated-idiom). The main effect of prime type emerged from GYYs (Wald χ^2 = 9.606, df=2, p=.008); the related-novel condition patterned with the related-idiom condition (p=.487), and the unrelated-idiom condition generated significantly more judgment errors than the related-novel (p=.004) and related-idiom conditions (p=.025). However, the main effect of prime type remained absent for CYs (Wald χ^2 = .955, df=2, p=.620) with an insignificant result found for both contrasting across the three conditions. It can thus be concluded that the absence of priming effects found in CYs' judgement data is not a result of potential flooring effects.

Discussion

To summarize the results, the internal lexical access has been observed for GYYs but not for CYs during the processing of whole idiom forms.

For GYYs, both RTs and accuracy showed the same pattern, as both the related novel phrases and related GYYs showed significant priming effects relative to the unrelated idioms, and no difference was found between the related novel phrases and the related GYYs. These results first indicate that the second constituent words were both literally activated and provided semantic cues for the related target DSWs. The lack of significance observed between novel phrases and GYYs further suggests that GYYs may have the same lexical nature as novel

phrases. Regarding the lexical identity of GYYs, several proposals have been put forward. Zhou (1998) argued that three-syllable GYYs like 笑面虎 smile-face-tiger "a smiling tiger - an outwardly kind but inwardly cruel person" should be categorized under the same group of three-syllable words like 明信片 name-mail-card 'postcard'. He claimed that in quantity, most three-syllable configurations in Chinese are words, and so it is natural to group three-syllable GYYs into the word category even though the semantic nature of 笑面虎 smile-face-tiger is semantically opaque and 明信片 name-mail-card is semantically transparent. The current study contributes new evidence on the wordhood of GYYs from the perspective of language processing. Based on the priming patterns observed for GYYs and matched phrases, we argue that GYYs are cognitively categorized as phrases by native speakers.

For CYs, RT data returned significant priming effects only for the related novel phrases. The related and unrelated CYs failed to provide a semantic cue for the target DSWs. The results show that the second constituent words in related CYs were not activated. This result may be interpreted under a different framework of idiom processing. According to the lexical representation hypothesis, CYs may be directly retrieved together with their figurative meanings from the lexicon. Therefore, within the brief time period employed (e.g., an SOA of 400 ms), no subsequent activation of the literal interpretation of CYs could be achieved. As an alternative the configuration view of interpretation would be that the lack of evidence of internal literal activation attributed to the fact that CYs are identified as idioms as soon as the initial encounter, and thus, the process of literal meaning configuration is stopped promptly at a very early stage. Then, if the findings in Holsinger (2013) hold true for Chinese, we can assume that lexical activation should be observed when the word of interest is the very first constituent word of an

idiom just as the word *foot* (literally related to the word *kick*) was found to draw a significant amount of attention from participants when it was prompted by idiom *kick the bucket*.

Regarding the lexical nature of CYs, the findings of the current study show that CYs are different from novel phrases. However, a further assumption may not be made regarding the wordhood of CYs even from a psycholinguistic point of view unless the same priming pattern was found in the difference between related CYs and matched related four-syllable words. However, the wordhood of a four-syllable configuration (e.g., 直升飞机 direct-rise-fly-machine 'helicopter') remains arguable. This is likely why Zhou (1998) suggested that four-word CYs should be considered phrases because most four-syllable units in Chinese are not words.

On the other hand, CYs do have perceptual salience compared to GYYs in Chinese. Just as Chen and Shu (2001) and Hoosain (1992) claimed for two-character words, because two-character words have perceptual salience in Chinese, two-character words are not processed as two separate units. Following this logic, a four-word CY may also be perceived as an inseparable chunk because of its salience without being completely lexicalized but having become frozen enough to be recognized as one unit.

In comparing the two types idioms, the opposite priming patterns were observed. Because the primary focus of this study is to investigate lexical access in idiom processing, regarding the contrast between two types of idioms, as we only intend to qualitatively compare priming effect patterns obtained for the two types of idioms, the lexical properties and second constituent words were not matched across GYYs and CYs. Nevertheless, analyses were conducted to compare mean values for meaningfulness, familiarity, and compositionality for the selected GYYs and CYs to develop an impression of how the two groups of idiom stimuli differ or resemble one another. Table 4.5 displays descriptive statistics and results for independent t-tests on average

ratings based on the norms in Study 1 on the five linguistic dimensions of meaningfulness, familiarity, compositionality, literality, and predictability for the selected GYYs and CYs.

Table 4.5: Descriptive statistics and independent *t*-test results on average ratings for the selected GYYs and CYs on five linguistic dimensions

	MEA		**FAM**		**COM**		**LIT**		**PRE**	
	CY	GYY	CY	GYY	CY	GYY	CY	GYY	CY	GYY
Mean	4.79	4.82	4.62	4.44	3.86	3.49	3.71	3.62	.93	.80
SD	.12	.11	.20	.24	.46	.58	.414	.45	.112	.24
t-value	-1.05		3.17		2.76		.81		2.67	
p-value	0.30		.00		.01		.42		.01	

MEA, meaningfulness; FAM, familiarity; COM, compositionality; LIT, literality; RRE, predictability

The independent t-test results show that the two idiom types are not significantly different on the dimension of meaningfulness and literality, but significantly different on familiarity, compositionality, and predictability. indicating that speakers understand the two types of idioms equally well. Ratings on familiarity are significantly higher for CYs than GYYs, indicating that the selected CYs are more frequent than GYYs. Ratings given on compositionality and predictability are significantly higher for CYs than they are GYYs, indicating that the selected CYs are more compositional and predictable than the selected GGYs. The statistics suggest that the CYs are more frequent, more compositional, and more predictable than GYYs.

Combining the statistics and with the priming patterns of the experiment, it can be seen that the more compositional idioms, CYs, did not provide any semantic cues for the target words while the less compositional ones, GYYs, did. This result runs counter to the prediction that more compositional idioms are more likely to be analyzed because their meanings receive more direct contributions from their constituents (cf. Gibbs & Gonzales, 1985). However, our finding, on the other hand, is not surprisingly new. Gibbs, Nayak, and Cutting (1989) also found that non-compositional idioms, which were predicted to be understood more readily because their forms

are more lexicalized, were actually processed slower than the compositional ones. Similarly, Burts (1992) reported that semantic transparent idioms were responded to faster than the semantically opaque idioms. Libben and Titone (2008) obtained facilitative effects of composability only for tasks requiring participants to overtly judge semantics. Even under an overt semantic judgment paradigm, Tabossi, Fanari, and Wolf (2009) found that participants were equally fast in judging decomposable and non-decomposable idioms as their matched controls. These findings indicate that composability plays a limited role in automatic processing. Moreover, the contradiction between native speakers' ratings on compositionality and the automatic processing found by Libben and Titone (2008) and in the present study may be indicative that native speakers' intuitions on the relationship between constituents and whole forms may be used to denote the lexical properties of idioms but are not sufficient enough to reflect native speakers' cognitive processes. As claimed by Tabossi, Fanari, and Wolf (2008), native speakers' intuitions on the semantic compositionality of idioms can be unclear and inconsistent. Tabossi, Fanari, and Wolf (2009) suggested that it is familiarity rather than compositionality that explains the rapid recognition of idioms. This assumption is compatible with our findings. Because CYs are more familiar to speakers than GYYs, CYs were identified faster than GYYs. As the configuration hypothesis predicts, as soon as an idiom is identified as an idiom, literal activation stops (Cutting & Bock, 1994). Therefore, the literal activation of a CY was inhibited sooner than that of a GYY. Furthermore, Chinese CYs are highly identifiable idioms. Unlike many idiomatic strings in other languages that can be completed literally until its last word (Cacciari & Tabossi, 1988), high-frequency CYs can be recognized by native speakers of Chinese after merely being presented with the first two words. According to Titone and Connine's (1994) findings, no literal activation may be observed in later stages of processing if

idioms are being highly predictable. Thus, based on our current findings, we speculate that for highly predictable idioms such as CYs, literal activation halts at a very early stage.

Limitations and Future Directions

This study proposes a working hypothesis regarding the different lexical status of GYYs and CYs. However, at this point, the findings cannot lead to a conclusive assumption because there are several concerns to be addressed about the results of the present experiment. The first is the possibility that the equal SOA was used for both GYYs and CYs. It is possible that because GYYs are one word shorter than CYs, the same SOA allows speakers to do closer scrutiny into the GYYs than CYs, which results in internal lexical activation of GYYs. Although it is possible that a short SOA makes the analysis of CYs difficult, the same difficulty also applies to the processing of four-word novel phrases. However, lexical activations were observed for the four-word novel phrases despite the same 400 ms SOA. Therefore, SOA may not be responsible for the difference between GYYs and CYs. Nevertheless, further studies that use adaptive SOAs (cf. Cutting & Bock, 1997) or contrast short, medium, and long SOAs (cf. Neely, 1976, 1977) can provide a more conclusive account for the different lexical activation observed.

Another concern relates to the syntactic structures of the GYYs. The selected CYs are mostly verbal (except two items 青梅竹马 *qing-mei-zhu-ma* green-plum-bamboo-horse “childhood sweetheart” and 风花雪月 *feng-hua-xue-yue* wind-flower-snow-moon “romantic themes”). However, GYYs consist of half verbal items (e.g., 敲竹杠 knock-bamboo-lever “take advantage of someone’s being in a weak position”) and half nominal items (e.g., 飞毛腿 fly-feather-leg “fleet-footed runner”). Syntactic structures have been proposed to be active and influential during idiom processing (Cutting & Bock, 1997). Therefore, further studies need to control the

verbal and nominal syntactic structures in order to confirm whether syntax plays any role in the semantic activation of the internal words. At this point, a post hoc analysis was conducted for verbal GYYs and nominal GYYs, respectively. The same priming effect pattern was observed for both structures, that related novel = related idiom > unrelated idiom, but the priming effects were only marginal (p=.06) for the nominal GYYs. This finding could be suggestive of the need to distinguish the two GYYs. As Su (2008) suggested, Chinese GYYs with a modification-modified structure such as 飞毛腿 demonstrate stronger lexicalized tendencies than GYYs with a verb-object structure such as 敲竹杠, as 93% of GYYs with a modification-modified structure are nominal, and a frozen nominal form is more likely to be identified as a word.

Another concern of the experiment design is the use of traditional characters in target word presentations. Although the participants did not report to have encountered significant difficulties in reading traditional characters, some traditional characters are admittedly more complex to recognized. To examine whether orthographic complexity plays a role in the current study, post hoc analyses were conducted using GLMMs with the inversely transformed RTs set as the dependent variable, prime condition (related idiom vs. related novel vs. unrelated idiom) and orthographic complexity of the target disyllabic words (zero traditional characters vs. one traditional character vs. two traditional characters) as fixed factors, and subjects and items as random factors. For CYs, results showed the main effect for prime condition, $F(2, 1618)=4.383$, $p=.013$, and orthographic complexity, $F(2, 1618)=74.986$, $p<.000$. Pairwise analyses revealed the same priming effect pattern for the three prime conditions. The related novel condition ($Mean_{-Raw\ RT}$=997 ms) showed greater priming effects than the related idiom condition ($Mean_{-Raw\ RT}$=1034 ms), Estimate=.039, SE=.016, t=2.453, p=.014, and than the unrelated idiom condition ($Mean_{-Raw\ RT}$=1035 ms), Estimate=.042, SE=.016, t=2.658, p=.008. The related idiom condition and the

unrelated condition showed no significant difference, Estimate=.003, SE=.016, *t*=.211, *p*=.833. However, the judgment data did not return the main effect for prime condition, *F*(2, 1795)=1.793, *p*=.167 when the orthographic complexity (*F*(2, 1795)=17.591, p<.000) was added as a fixed factor. These patterns were same as the patterns without considering orthographic complexity. For GYYs, RT data returned the main effect for prime condition, *F*(2, 1649)=5.760, *p*=.003, and for orthographic complexity, *F*(2, 1649)=48.357, *p*<.000. Pairwise contrasts revealed same priming effect patterns as before, with the related novel ($Mean_{\text{-Raw RT}}$=933 ms) and related idiom conditions ($Mean_{\text{-Raw RT}}$=943 ms) showing no significant difference (Estimate=-.011, SE=.018, *t*=-.625, *p*=.532), but the unrelated idiom condition ($Mean_{\text{-Raw RT}}$=1000 ms) being significantly slower than the related novel (Estimate=.057, SE=.018, *t*=3.206, *p*=.001) and related idiom conditions (Estimate=.046, SE=.018, *t*=2.602, *p*=.009). The judgment data returned the main effect for orthographic complexity, *F*(2, 1795)=12.939, *p*<.000, but the main effect for the prime condition was marginal, *F*(2, 1795)=2.932, *p*=.054. Pairwise analyses revealed a significant difference between the related novel and unrelated idiom conditions, Estimate=-.034, SE=.015, *t*=-2.240, *p*=.025. The difference between the related idiom and unrelated idiom conditions was marginal, Estimate=-.027, SE=.016, *t*=-1.717, *p*=.086. The difference between the related novel and the related idiom conditions was not significant, Estimate=-.008, SE=.014, *t*=-.551, *p*=.582.

To summarize, the priming effect patterns basically hold for both GYYs and CYs when the orthographic complexity was added as a fixed factor. The reason is that the orthographic complexity is consistent across three conditions. However, we did find that the orthographic complexity had significant influences on both RTs and judgment accuracy. Future research should more appropriately control the orthographic factor in order to have more robust results.

CHAPTER 5: STUDY 3

THE PROCESSING OF IDIOMATIC AND NON-IDIOMATIC LEXICAL BUNDLES BY L1 AND L2 SPEAKERS: WHAT CAN THINK-ALOUD PROTOCOLS TELL US?

Although mounting evidence has reported that high-frequency lexical bundles (LBs) have processing advantages over matched novel phrases, few studies have dedicated to compare if different types of LBs are processed differently. Besides, little was known about what native speakers (NSs) and nonnative speakers (NNSs) are thinking while comprehending different LBs and how their thought processes and processing strategies are different from one another. This study aims to answer these questions by triangulating dichotomous judgements, response times (RTs), and think-aloud (TA) protocols. Forty NSs and advanced NNSs read idioms (3- and 4-characters) and matched non-idiomatic formulaic sequences (FSs) in a silent grammaticality judgment task (GJT) and a TA GJT with a one-week interval. RT data showed NSs are more sensitive to stimuli type while NSSs are more sensitive to stimuli length, suggesting idioms and FSs have different lexical representations in NSs' but not NNSs' lexicons. TA verbalizations showed that NNSs used more analytical strategies to comprehend idioms and their knowledge of idioms was partially correct or incorrect even though they were able to make correct Yes-or-No judgments on those idioms. TA data contribute insights into learners' cognitive processes and highlight potential methodological issues in the LB research in the second language.

Introduction

Lexical bundle, interchangeably used with the term formulaic language, is a sequence of words that frequently recur and are used as a whole unit by NSs (Biber & Conrad, 1999; Bybee

& Hopper, 2001; Wray, 2000). In recent decades, how second language (L2) learners acquire and comprehend LBs has been one of the fast-growing areas of research in the second language acquisition (SLA) field. The increasing attention is based on the assumption that the acquisition of LBs is able to facilitate the overall language proficiency (N. Ellis, 2012; Weinert, 1995; Wray, 2000; Schmitt, 2012; Wood, 2006). This assumption is based on the accumulating empirical findings from different angles. Corpus-based research has found that a small class of LBs covers a relatively large portion of spoken and written texts (Oppenheim, 2000; Moon, 1998; Sorhus, 1977; Erman & Warren, 2000). The ubiquity proves the importance of LBs in language use (Schmitt, Dörnyei, Adolphs, & Durow, 2004). Research examining LB's lexical properties showed the semantic non-decomposability and integrity of LBs (Swinney & Cutler, 1979; Gibbs, 1980) allows a relatively complicated meaning and function to be integrated in relatively shorter sequences. The compact structure of LB can lead to the communication efficiency (Nattinger & DeCarrico, 1992, p. 62–63). Research focusing on LB processing has found that first language (L1) speakers are able to retrieve LBs as whole units directly from long-term memory (Pawley & Syder, 1983), and LBs have a processing advantage over rule-generated phrases (Carrol & Conklin, 2014; Schmitt & Underwood, 2004; Nekrasova, 2009; Jiang & Nekrasova, 2007). All these findings favor the argument that the mastery of LBs could lead to an increase in learners' language proficiency (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Gardner & Davies, 2007; Lindstromberg & Boers, 2008; Rott, 2009). However, only a few studies (Boers & Demecheleer, 2001; Spöttl, & McCarthy, 2004; Irujo, 1986; Cooper, 1999; Myles, Hooper & Mitchell, 1998) have focused on describing the quality of LB knowledge that L2 learners possess. Do L2 learners fully understand the meaning and usage of an LB? What kinds of strategies do L2 learners use to understand the meanings of LBs? The reason that these issues

have not been comprehensively addressed may be due to the research orientations and the methodologies used.

First, research on LB processing in L2 has mostly concentrated on discussing different processing models rather than observing the nuanced manifestations of learners' behavior to identify the quality and depth of learners' knowledge of LBs. For example, studies gathering learners' metalinguistic judgments/ratings (e.g., Abel, 2003; Tabossi, Wolf & Koterle, 2009) response times (RTs) (e.g., Jiang & Nekrasova, 2007; Millar, 2010; Conklin & Schmitt, 2008) and tracking eye movements (e.g., Underwood, Schmitt, & Galpin, 2004; Siyanova-Chanturia, Conklin, & Schmitt, 2011) aim to examine whether LBs are processed as whole units, whether the literal meanings of the constituent words are accessed, or whether both lexical access and whole concept retrieve are active simultaneously, which is the main debate among the lexical representation hypothesis (Swinney & Cutler, 1979), the configuration hypothesis (Cacciari & Tabossi, 1988), and the hybrid view (Cutting & Bock, 1997; Abel, 2003). Studies analyzing the correlations between learners' performance and LB's linguistic properties of LBs (e.g., N. Ellis, Simpson-Vlach, & Maynard, 2008; Nekrasova, 2009) generally aim to explore what factors (e.g., frequency or mutual information score) predetermine the psychological reality of LBs, which is the central argument of usage-based approaches (Bybee & Hopper; 2001; N. Ellis, 2002; Weinert, 2010). These theory-oriented studies have made significant contributions to the definition, classification, and characterization of LBs from learner-internal perspectives. If the purpose of LB research extends to providing more direct insights to L2 learning, the quality of learners' knowledge of LBs merit a scrutiny. As Myles and Cordier (2017) illustrated, if the spoken formula "you know what I mean" is produced "haltingly or with errors, e.g., 'you…uhm…know…what uhm mean'" this formulaic sequence (FSs) certainly has become an

internalized knowledge to the L2 speaker. This example demonstrates the necessity to gather the data that are able to reflect the quality of learners' knowledge during the online LB processing. Nevertheless, research on LB comprehension has concentrated mainly on quantitatively measurable data such as event-related potentials, RTs, eye movements, dichotomous judgments, and controlled productions. SLA research should involve both "measuring and describing learners' knowledge of a language" (Bowles, 2010a; p. 1). Just as Read and Nation (2004) argued, "an adequate account of formulaic units as they function in language acquisition and language use can come only from a combination of quantitative and qualitative analyses." Moreover, the qualitative data mentioned above are primarily used to measure native speakers' (NSs) language processing. Gass (1983, p. 273) pointed out if we assume that NNSs' language is similar to NSs' language, then it is reasonable to assume that both languages can be measured by the same means. However, learners' knowledge can be incomplete or indeterminate, and it is thus important to know what data are "truly representative of a learner's knowledge and what [are] not" (Gass, 1994; p. 305). Therefore, data used to measure native language processing may not be sufficient to reflect learners' processing of the second languages (R. Ellis, 1991; Myles & Cordier; 2017). To obtain more valid results and fuller picture about learners' knowledge of an L2, Leow, Grey, Marijuan, and Moorman (2014) suggested that multiple types of data need to be compared. The present study re-examined the nonnative speakers' (NNSs) processing of LBs through triangulating three qualitative and qualitative data, RTs, dichotomous judgments, and TA verbalizations elicited from two GJTs. One goal is to measure and describe NNSs' knowledge of LBs and see if learners' verbalizations on how they understand LBs can tell us something that RTs and dichotomous judgments cannot. The other goal was to compare the processing of two types of LBs, namely, idioms, such as "kick the bucket", and non-idiomatic

rule-generated formulaic sequences, such as “you know what I mean”. Idioms are prototypical LBs, while FSs are considered marginal members of the LB family (Schmitt & Carter, 2004). Whereas evidence has been found that the highly familiar non-idiomatic everyday phrases have the same processing advantage as idioms do (Glass, 1983; Burt, 1991), the syntactic and semantic differences between idioms and FSs seem to suggest that they should be treated differently in L2 teaching. For example, idiomatic expressions (e.g., 不敢当 *bu-gan-dang* not-dare-be, “I really don’t deserve this compliment”) are often included in the vocabulary list in L2 textbooks and taught as single words, while the equally frequent FSs (e.g., 不敢动 *bu-gan-dong* not-dare-move, “dare not to move”) are not. Does this treatment make sense? Do L2 learners also comprehend the two forms differently? The study also set out to address these questions.

Literature Review

The Processing of different types of LBs by L2 learners

As Carrol and Conklin (2019) suggested, studies of individual types of LBs art attributed to the contributions of specific linguistic factors, but little work has been done to compare how different types of LBs with different properties are processed. SLA-oriented LB processing studies also have called attention to the necessity of distinguishing different types of LB according to their different syntactic-semantic properties and functions (Conrad & Biber, 2005; Wray, 2004; Myles & Cordier, 2017). These studies suggested that mixing different types of LBs in one study may weaken the validity of the results. Moreover, different types of LBs are acquired through different sources. For example, idiom knowledge is more likely obtained from formal instructions in classrooms, while some spoken formulae can be learned from everyday communications in the target language environment. The source of L2 knowledge has been

found to have an impact on the degree of mastery of LBs (Yamashita & Jiang, 2010; Meunier, 2012). Therefore, it is necessary for SLA-oriented LB research to refine the classification of the research targets. Some research has already been conducted in relation to this endeavor. Nekrasova (2009) conducted a gap-filling task and a dictation task to compare L1 and L2 English speakers' knowledge of discourse-organizing bundles and referential bundles. The results showed that both NSs and NNSs knew more discourse-organizing bundles than referential bundles. Participants' self-reports also supported the argument that because the discourse organizers connected larger portions of texts, they could facilitate the overall comprehension of the topic (Biber, Conrad, & Cortes, 2004). Using eye-tracking measurements, Siyanova-Chanturia, Conklin and Schmitt (2011) compared the processing of idioms' figurative meaning (e.g., at the end of the day – "eventually") to that of their literal meaning (e.g., at the end of the day – "in the evening"). NSs showed no processing advantage for the figurative uses over the literal uses. NNSs processed idioms with figurative uses more slowly than idioms with literal uses. Another important contribution is that learners' eye movements revealed that there is an idiom key within each idiom that slows the processing of the figurative meaning. In summary, idiomatic processing did not show an advantage over non-idiomatic processing for NSs and NNSs. Irujo (1986) distinguished three different types of English idioms based on their semantic similarity with the participants' L1, Spanish, to investigate L1 transfer in the learning of idioms in an L2. The results showed that idioms with identical meanings in two languages were the easiest to comprehend and produce. Similar idioms were comprehended almost as well but showed interference from Spanish. Different idioms were the most difficult to comprehend and produce but showed less interference than similar idioms. Within each type, the idioms that had transparent structures, simple constituent words, and high frequencies were comprehended and

produced most correctly. Similarly, in a study on the influence of L1 on the acquisition of L2 collocations, Yamashita and Jiang (2010) utilized a phrase-acceptability judgment task to compare Japanese L1 English-as-a-second-language (ESL) learners and English-as-a-foreign-language (EFL) learners' processing of two types of collocations. One was congruent collocations, whose lexical components are similar in L1 and L2, and the other was incongruent collocations, whose lexical components differ in the two languages. The results showed that EFL learners made more judgment errors with and responded more slowly to incongruent collocations than to congruent collocations. ESL users generally performed better than EFL learners, but they still made more errors on incongruent collocations. The findings suggested that the acquisition of L2 collocations is a long-term process in which both L1 congruency and L2 input have interactive effects. A methodological difference between the two studies mentioned above is that Irujo utilized a diversity of tasks (idiom-meaning association, idiom definition, discourse completion, and idiom translation) and collected both quantitative data and qualitative data. The use of two types of data allow researchers to generalize how much learners know and do not know and the depth of learners' knowledge about the target LBs.

Think-aloud protocols in SLA research

TA protocols represent a concurrent data collection procedure that has been widely used to measure learners' lexical knowledge in SLA fields (e.g., Bowles, 2004; Lawson & Hogben, 1996; Spöttl, & McCarthy, 2004; Nassaji, 2006; Read, 1993; Qian, 1999; Haastrup & Henriksen, 2000). Viewing as a channel for examining the depth of processing (e.g., Fraser, 1999; Yanguas, 2009; Morgan-Short, Heil, Botero Moriarty & Ebert, 2012; Adrada-Rafael, 2017), the amount of attention (e.g., Leow, 1997, 1998, 2000, 2001; Rosa & Leow, 2004a, 2004b; Rosa & O'Neill, 1999), and the strategies employed by learners when processing L2 input (e.g., Whalen &

Menard, 1995; Cohen, 2000). L2 researchers use TA protocols to gather data about learners' thought processes for a variety of theoretical and applied purposes (Bowles, 2010a, 2010b; Leow, Hsieh & Moreno, 2008; Leow, Grey, Marijuan & Moorman, 2014). Bowles (2010) provided an overview of research using TA data and meta-analytic research and found this procedure to be valid if implemented appropriately. A typical TA procedure starts with one or two sentences that briefly reiterate why the participants are being asked to think aloud without giving away any information about the goal of the study. An example is as follows: "In this experiment, I am interested in what you think about when you complete these tasks. In order to find out, I am going to ask you to THINK ALOUD as you work through the mazes (Bowles, 2008)." Following the rationale instruction, more specific instruction on how to think aloud while performing a task must be provided, including what thinking aloud involves, what language they should use to verbalize their thoughts, and the level of detail required in thinking aloud. After the participants signal that they understand the instructions, some practice trials are usually conducted to familiarize the participants with either the TA procedure or the task they will perform. Finally, in the real trial, to ensure validity, one researcher should be present with the participant and remind him/her to think aloud whenever he/she engages in silent thinking.

One methodological controversy of think-aloud protocols that TA research must address is reactivity. Reactivity is the potential that the act of thinking aloud may alter subjects' cognitive processes while performing a task. However, both meta-analysis study (Bowles, 2010) and empirical research (Leow & Morgan-Short, 2004; Bowles & Leow 2005) have shown that the TA procedure does not have significant detrimental or facilitative effects on L2 learners' performance. The authors also suggested that any research employing TA procedure should

include a control group who does not think-aloud to ensure no detrimental effect of the act of thinking-aloud on subjects' performances.

As a versatile tool, TA protocols can also be used to complement the findings of other concurrent data collection procedures in studies that aim to elicit both qualitative and quantitative data on learners' performance (Mackey, Gass, & McDonough, 2000; Leow, Grey, Marijuan & Moorman, 2014). A number of SLA studies have compared TA approaches to other data-gathering techniques or language assessment measures. The purpose of such research is either to validate/question a particular research methodology or obtain information on the participants' thought processes, strategies or depth of knowledge that single measures or other instruments may not be able to provide. Godfroid and Schmidtke (2013) used a combination of TA protocols, eye-tracking, and posttests to investigate incidental vocabulary learning. Rebuschat, Hamrick, Riestenberg, Sachs, and Ziegler (2015) conducted triangulating TA protocols, retrospective verbal reports, and other subjective assessments, such as confidence ratings, to investigate how NNSs allocate awareness under different incidental learning conditions. Nassaji (2006) compared learners' reading comprehension in a TA condition and a silent reading condition to investigate ESL learners' depth of vocabulary knowledge and lexical inferencing strategies in deriving word meaning from context. Rosa and O'Neill (1999) utilized a recognition test and TA protocols to investigate how learner intake may be affected by the allocation of awareness and by different levels of explicit presentation. Morgan-Short, Heil, Botero-Moriarty, and Ebert (2012) compared TA reading and traditional silent reading comprehension results to discuss whether attending to grammatical or lexical forms while reading for meaning affects the comprehension of the text; they found that learners who attended to the lexical and grammatical forms exhibited greater evidence of comprehension. Targeting the

guesswork problem, Kamimoto (2008) added a TA procedure to a recognition task, and the results of the TA verbalizations were found to be more precise in reflecting learners' actual vocabulary knowledge. In a discussion of the reliability of grammaticality judgments in L2 studies, R Ellis (1991) asked Chinese learners of English to think aloud when retaking an untimed grammaticality test that the participants had taken one week before and found that learners were inconsistent in 22.5% of their judgments.

Following the comparative approaches, the goal of the present study was to triangulate three data sources, RTs, dichotomous (Yes/No) judgments, and TA verbalizations to obtain a more comprehensive picture of picture of how well learners understand idioms and FSs. In the present study, FSs were defined as non-idiomatic rule-generated phrases that are often considered fixed expressions (cf. Jiang & Nekrasova, 2007). The following three research questions (RQs) were addressed:

1. What do RT data reveal about NSs and NNSs' knowledge of idioms and FSs?
2. What do dichotomous judgments reveal about NSs and NNSs' knowledge of idioms and FSs?
    (a) Do NSs and NNSs judge the same stimuli in the two GJTs with the same degree of accuracy?
    (b) Do NSs and NNSs judge the same stimuli consistently in the two GJTs?
3. What do TA verbalizations reveal about NSs and NNSs' knowledge of idioms and FSs?
    (a) Can dichotomous judgments reflect NSs and NNSs' actual knowledge of idioms and FSs?
    (b) Do NSs and NNSs use the same strategies to process idioms and FSs?
    (c) Is processing strategy correlated with the accuracy of dichotomous judgments?

Method

Participants

The NNS participants were 23 Chinese language learners (13 females; 10 males) recruited from four Chinese universities in Beijing. They were originally from nine countries, including Egypt, South Korea, Thailand, Kazakhstan, Outer Mongolia, Russia, Vietnam, Japan, and Nepal, with a mean age of 22.5 years. All the NNS participants had passed the ***H****anyu* ***S****huiping* ***K****aoshi* 6 (Chinese Proficiency Test 6; hereafter referred to as the HSK) within the last two years, with an average score of 210.09 (full score=300). The HSK is a standardized Chinese proficiency test that assesses nonnative Chinese speakers' ability to use Chinese in daily life and in academic and professional settings. The latest version of the HSK consists of 6 levels. The HSK 6 is the highest level and is intended for advanced learners and requires a vocabulary size of a minimum of 5,000 words. Twenty Chinese NSs (12 females; 8 males; M_{age}=27.5) were recruited from two Chinese universities as the control group.

Test materials

Thirty-six Chinese as a second language (CSL) teachers with ten years or more of teaching experience were asked to rate the likelihood of an HSK 6 learner knowing the preselected idioms using a 1~5 Likert scale ranging from "Must know" to "Impossible to know". Idioms that received an average rating higher than 4 were included in the test material. For example, 80% of the teacher raters identified idiom 不约而同 *bu-yue-er-tong* not-arrange-and-same "do or think the same without prior consultation" as a "Must know" idiom; it received an average score of 4.5 and was therefore included in the test. In contrast, idiom 打官腔 *da-guan-qiang* play-official-tone "speak in a bureaucratic tone" received an average score of 2, which is

equivalent to “Somewhat difficult to know”; therefore, this candidate idiom was excluded. Finally, twenty-four 3-character idioms (3-idioms) and twenty-four 4-character idioms (4-idioms) were selected for the test material. Each idiom (e.g., 走后门 *zou-hou-men* walk-back-door “to secure advantage by some under-the-table means”) was then matched with an FS (e.g., 走出去 *zou-chu-qu* walk-out-go “walk out”) containing at least one content word identical with its idiom counterpart. Efforts were also made to ensure each idiom-FS pair shared as many semantic and syntactic elements as possible. The whole form’s log frequency[11] (Log(f)) and stroke were matched across the three-character sequences and the four-character sequences. Descriptive statistics and two sample t-test results are presented in Table 5.1. Appendix E presents a full list of the test materials.

Table 5.1: Descriptive statistics and t-test results of the selected idiom-FS pairs

	3-idiom	3-FSs	4-Idioms	4-FSs
Mean of Log(f)	3.27	3.02	3.64	3.58
Mean of Strokes	21.25	20.85	26.25	25.16
t(23)	1.53		1.67	
p	1.33		.11	

In summary, four types of LBs were included in the study: 3-idioms, 4-idioms, three-character FSs (3-FSs), and four-character FSs (4-FSs). Another 96 ungrammatical phrases were included as filler items. All test items were evenly divided into two counterbalanced blocks (A and B), and each block consisted of twenty-four 3-idioms, twenty-four 4-idioms, twenty-four 3-FSs, twenty-four 4-FSs, and forty-eight ungrammatical phrases. A block that contained an idiom did not also contain the matched FS.

[11] The raw frequency was based on the token frequency taken from BBC corpus (BLCU Corpus Center; Xun, Rao, Xiao, & Zang, 2016), amounting to 15 billion characters.

Procedure

Character quiz

Within one week before the main test, the L2 learner participants were given a character list to study at home. The list included (but was not limited to) all the characters that would appear in the test material. All the characters were within the required vocabulary range of the HSK 1 to 4 guidelines, so presumably an HSK 6 learner would already know all of the characters. Before the day of the first session, all the participants took a computerized character quiz in which they were asked to associate each character with its correct meaning in a multiple-choice task. Two participants did not get 100% correct on the quiz. They were still invited to the following sessions of the study, but their data were excluded from the analyses.

Main test

All participants performed in the two GJTs[12] on two different days with a one-week interval. Both the NSs and NNSs were further divided into two groups (Group 1 and Group 2) with approximately the same number of subjects in each group. To examine the reactivity effect, Group 1 and Group 2 performed the two GJTs in a counterbalanced order. For each GJT session, participants saw both the blocks (A and B) of the test material with a 5-minute rest period in between. The two blocks of material were repeatedly used in the two GJT sessions but presented to the same participant in a different order. Figure 1 demonstrates the detailed procedure.

[12] In the literature, the terms grammaticality judgment and acceptability judgment are used interchangeably, and both can indicate whether or not stimuli are likely to be found or acceptable in a given language (Gass, 1994). In this study, we followed Ellis, Simpson-Vlach, Maynard (2008) and used the term "grammaticality".

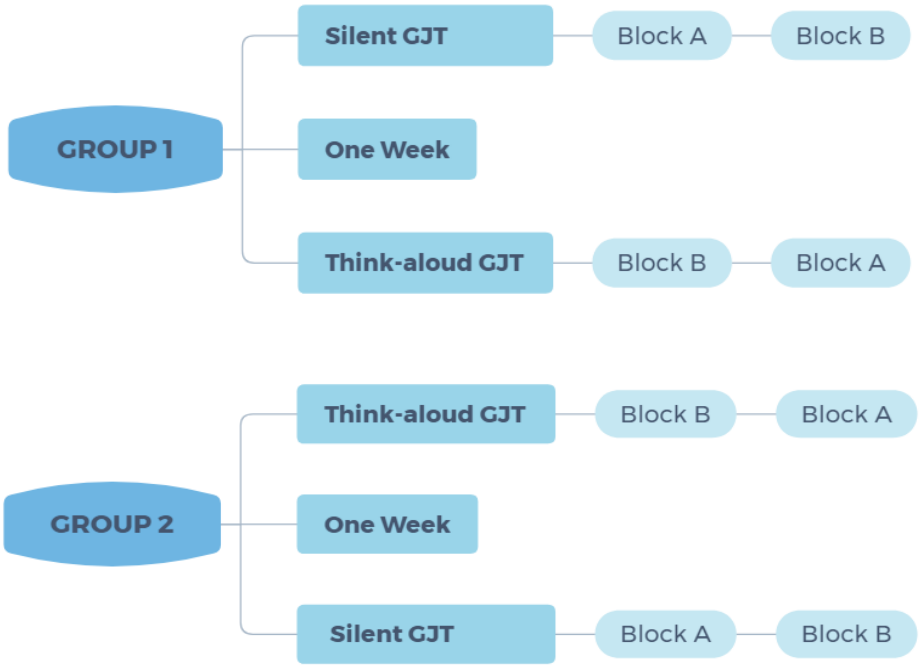

Figure 5.1: Experimental procedure

SILENT GJT In the silent GJT session, the participants had a brief instruction session followed by a ten-trial training. In the experiment, a fixation cross appeared in the center of the screen for 800 ms and then disappeared. Then, an LB was presented in the same position. The participants were asked to judge whether the LBs they saw on the computer screen were likely to be found in Chinese by pressing a corresponding key (A for "Yes" and L for "No"). The participants were told that they could take as long as they needed to make a decision. The LB remained on the screen until the participant pressed a response key. After each judgment, the participants pressed the spacebar to proceed to the next trial. After finishing the first block, the participants took a 5-minute break and then returned to the second block. The experiment was run on Lenovo computers in Microsoft Windows 11 programed by Paradigm.

TA GJT In the TA GJT session, the participants were instructed to think aloud while completing the GJT. All stimuli were presented on a computer screen one after another by the

participants pressing the spacebar. Before the experiment, written instructions were provided to the participants to ensure that they understood they should verbalize whatever thoughts went through their minds when performing the GJT. Next, the participants were 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 individuals understand language. They would therefore be asked to first read aloud the LB they saw on screen and then judge whether the LB was likely to be read or heard in Chinese while "externalizing your inner speech and speaking your thoughts aloud when you are making the judgment". The instructions emphasized the importance of the participants "speaking whatever can help you make a judgment on a stimulus" without worrying about giving explanations or using examples or incomplete sentences. Participants were asked to think aloud using the target language, Chinese. The examiner sat beside the participant and provided some prompts. To gain insight into exactly how much participants knew about the LBs and avoid disturbing the participants' flow of thought, two types of neutral prompts were given (van Someren, Barnard, and Sandberg, 1994). One involved asking "你在想什么 *ni zai xiang shenme*, 'What are you thinking about?'" when the subject fell into silent thinking for a long period, and the other was to ask "你是怎么知道的 *ni shi zenme zhidao de*, 'How do you know?'" when the subject provided only a Yes/No judgment without further verbalizing about the stimulus. One L2 learner responded, "I have no idea" for most of the items, and her data were excluded from further analysis. All TA verbalizations were audio-recorded with Audacity 2.3.0 and transcribed for coding and analysis.

Coding

Coding dichotomous judgments

Verbalizations collected in the TA session were first coded for participants' dichotomous (Yes/No) judgments on each item, as the participants were asked to provide a clear Yes/No answer before thinking aloud their thought processes. Their judgments were coded as "correct" or "incorrect". These data were compared against the button-pressing judgments elicited in the silent session.

Coding the status of knowledge

The validity of dichotomous judgments (Yes/No) has long been questioned in terms of measuring L2 competence because the binary coding tends to over- or underestimate the learner's lexical knowledge (Nation, 2001). N. Schmitt, D. Schmitt, and Clapham (2001) conducted a post lexical decision-task interview and identified three different levels regarding the status of lexical knowledge in L2 learners: no knowledge, partial knowledge, and full knowledge. Adapted from this categorization, we coded the TA data into four categories: (1) no evidence, (2) incorrect knowledge ("incorrect"), (3) partially correct knowledge ("partial"), and (4) fully correct knowledge ("correct"). Table 5.2 presents details about how the coding of the status of knowledge was operationalized.

Table 5.2: Operationalization of the status of knowledge

	No Evidence	**Incorrect**	**Partial**	**Correct**
Criteria	No evidence to tell whether or not participants understand the target.	Participants exhibit incorrect knowledge about the target.	Participants exhibit partial knowledge of the target.	Participants exhibit full knowledge of the target.
Evidence	• Acknowledge the target "has been heard or learned" but have forgotten the meaning. • Identify that the target is a certain type of expression. • Use "A just means A" as a justification.	• Judge a correct item to be incorrect because it has never been heard. • Provide a wrong interpretation or a wrong example. • Judge a correct item to be wrong and provide an alternative expression. • Provide an incorrect metalinguistic comment.	• Provide a literal interpretation of an idiom that has literal plausibility. • Provide a metalinguistic comment that does not demonstrate whether the participant understands the meaning of the target. • Provide a related but not precisely correct interpretation or scenario. • Provide an example sentence with correct but not precisely correct grammar. • Provide a correct sentence with a neutral context that does not ensure the participant knows the meaning.	• Provide a correct example sentence with a specific context. • Provide a correct interpretation of the target and a figurative interpretation if the target is an idiom. • Provide a correct metalinguistic comment that shows that the participant understands the meaning of the target.

Coding processing strategy

As Schmitt, Grandage, and Adolphs (2004) and Nekrasova (2009) suggested, there is no direct assessment as to whether or not lexical sequences are processed as holistic units. Thus, if participants thought processes could reveal in what way LBs are recognized and comprehended, it would be indicative of whether LBs are holistically retrieved or analyzed bit by bit. For this

purpose, the processing strategy reflected in the TA justifications was coded into two categories: holistic and analytical. A holistic strategy was assigned if the participant only mentioned, interpreted or explained the whole form. Evidence included a) providing intuitive comments (e.g., stating "I have seen/heard/learned this idiom" or "A just means A"), b) giving an example sentence (e.g., 考试前,他开夜车复习, "before the exam, he *kai-ye-che* reviewing", where *kai-ye-che* drive-night-car means "to stay up late working or studying"), and c) using another word to interpret the whole form (e.g., 要面子就是骄傲的意思 "*yao-mian-zi* just means '*jiao-ao*'", where *yao-mianzi* want-face is an idiom meaning "be keen on face saving"). An analytical strategy was assigned if the verbalization included mentioning, interpreting, or explaining the constituent character(s), word(s), or grammar(s). Evidence included a) giving a verbatim translation (e.g., 走后门就是从后面的门走 "*zou-hou-men* just means walking through the back door", where *zou-hou-men* walk-back-door can literally mean "to get in through the back door", and its figurative meaning is "to secure advantage by some under-the-table means"), b) providing an interpretation that mentioned some of the constituent words (e.g., 无能为力就是没有能力做什么 "*wu-neng-wei-li* just means not having the ability to do something", where the idiom *wu-neng-wei-li* no-capable-act-strength means "helpless, incapable of doing", and the subject mentioned three constituent words 无 *wu* "not to have", 能 *neng* "capable', and 力 *li* "strength" in his/her verbalization), and 3) giving a metalinguistic comment (e.g., 谈天说地是对的, 中文里有'什么天什么地'的结构 "*tan-tian-shuo-di* is correct; Chinese has the structure like "something-sky, something-earth", in which the idiom *tan-tian-shuo-di* talk-sky-speak-earth means "talk about everything under the sun", but the subject made a correct judgment based on the knowledge about the idiomatic frame "…sky…earth" but failed to give the exact meaning.).

Interrater reliability

The first author and a research assistant coded 25% of the data independently. Interrater reliability reached 100% for judgment and 94.6% for status of knowledge. These results were considered high enough for the first author to code the remaining 75% of the data alone.

Results

The results for the RQs of the study are reported in this section. Statistical analyses were performed using SPSS 25 with the default α level set at 0.05 (unless otherwise indicated) and the level of the confidence interval (CI) set to 95%.

Preliminary analyses

Before answering the RQs, preliminary analyses were conducted to examine the homogeneity of the four groups with test order as a condition. The purpose of these analyses was to ensure that any observed differences between the silent GJT and TA GJT were not a result of the act of thinking aloud, in other words, the reactivity effects. For that purpose, four independent samples *t*-tests (α level=0.0125) were performed with the dichotomous judgement scores and mean RTs as dependent variables and the test order (silent TA vs. TA silent) as the function for the L1 and L2 groups, respectively. The judgment accuracy results showed that the difference between the two L2 groups was nonsignificant ($t(18)=-0.788$, $p=.441$, CI=-10.26~4.66), with a small effect size, Cohen's $d=0.33$ (Plonsky & Oswald, 2014). Non-significance was also observed for the two L1 groups ($t(18)=0.416$, $p=.682$, CI=-1.62~2.42), with a small effect size ($d=0.15$). The RT data also did not yield significant differences for the two L2 groups ($t(18)=-1.384$, $p=.183$, CI=-1241.9~255.6, $d=0.61$) or for the two L1 groups ($t(18)=0.514$, $p=0.613$, CI=-140.5~231.6, $d=0.23$). These results indicate that test order does not

impact participants' performances. Therefore, four groups were therefore merged into two groups, with nativeness (L1 vs. L2) being the only between-group condition in the following analyses.

RQ1. What do RT data reveal about L1 and L2 speakers' processing of idioms and FSs?

RTs for incorrect responses that were three standard deviations from each participant's mean were excluded. This procedure eliminated 3% of the L1 data and 13.1% of the L2 data. The RT data were then transformed by the Lg10 algorithm to reduce skewness.

A generalized linear mixed model (GLMM) was performed with Group (L1 vs. L2), Type of stimuli (idiom vs. FS), Length of stimuli (3-character vs. 4-character) being fixed factors, Lg10(frequency) being covariate, and Subject and Item being random factors. Results are presented in Table 5.3. Lg10(RT) yielded the main effect of Group ($p<.000$), Type ($p=.012$), Length ($p<.000$), and of the interaction of Type × Group ($p<.000$), Length × Group ($p<.000$), and Type × Length ($p=.006$). The three-way interaction of Type × Length × Group is not significant ($p=.054$). Lg10(frequency) is a significant predictor to RTs ($p<.000$).

Table 5.3: Fixed coefficients for Lg10(RT)

	Coefficient	SE	*t*	CI
Group	-0.324	0.036	-9.123***	(-0.394, -0.255)
Type	-0.03	0.012	-2.503*	(-0.054, -0.007)
Length	-0.073	0.012	-5.883***	(-0.098, -0.049)
Group × Type	0.066	0.017	3.995***	(0.034, 0.098)
Group × Length	0.077	0.017	4.537**	(0.044, 0.110)
Type × Length	0.047	0.017	2.731***	(0.013, 0.080)
Group × Type × Length	-0.045	0.023	-1.926	(-0.091, 0.001)
Lg10 (frequency)	-0.025	0.007	-3.526***	(-0.038, -0.011)

To further examine the effect of Type and Length within each group, pairwise contrasts were computed with Bonferroni adjusted significance level set as .05. Results show that NSs are more sensitive to Type (p<.000) than to Length (p=.605), with idioms being responded to 107ms faster on average than FSs. Four-way pairwise contrast analyses further revealed that 3-idioms were responded to significantly faster than 3-FSs (p=.001), and 4-idioms significantly faster than 4-FSs (p=.002) by NSs. Differences between 3- and 4-idioms (p=.763) and between 3- and 4-FSs (p=.656) were not significant. In contrast, L2 learners are more sensitive to Length (p<.000) than Type (p=.419), with 4-character stimuli being 266ms slower on average than 3-character ones. Pairwise contrast analyses for the L2 group returned significant differences for 3- and 4-idiom pairs (p<.000), 3- and 4-FS pairs (p=.03), and 4-idiom and 4-FS pairs (p=.012), but not for 3-idiom and 3-FS pairs (p=.175). Figure 5.2 demonstrates average raw RTs of stimuli by Group, Type, and Length.

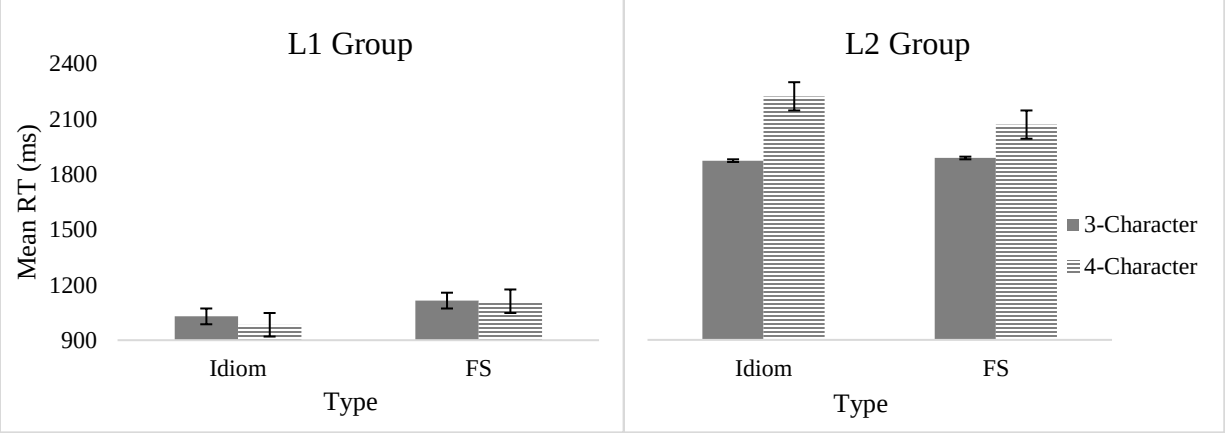

Figure 5.2: Average raw RTs by group, type, and length

It can be seen from Figure 5.2 that NSs recognize idioms much faster than non-idiomatic FSs regardless of whether they were 3-character or a 4-character in length, while NNSs recognize 3-character stimuli faster than 4-character stimuli regardless of their idiomaticity. In summary, the RT data revealed two different processing patterns for NSs and NNSs. NSs'

processing pattern was ranked as 4-idiom ≤ 3-idiom < 4 FS ≤ 3 FS ("≤" indicating nonsignificantly slower than), and L2 learners' processing pattern was ranked as 3-idiom ≤ 3-FS < 4-FS < 4-idiom.

RQ2. What do dichotomous judgments reveal about L1 and L2 speakers' knowledge of idioms and FSs?

To answer first sub-question of RQ2: ***do NSs and NNSs judge the same stimuli in the two GJTs with the same degree of accuracy***, GLMMs were constructed separately for the L1 group and the L2 group both with Session (silent vs. TA), Type of stimuli, and Length of stimuli as fixed factors, Lg10(frequency) as covariate, and Subject and Item as random factors. The results are presented in Table 5.4.

Table 5.4: Fixed coefficients for type, length, session and their interactions by group

		Coefficient	SE	*t*	CI	
L1 Group	Type	0.345	0.356	-0.967	(-1.044,	0.354)
	Length	-0.251	0.363	-0.692	(-0.962,	0.46)
	Session	0	0.386	0	(-0.756,	0.756)
	Lg10frequency	0.392	0.182	1.805	(-0.028,	0.686)
	Type × Length	0.486	0.495	0.983	(-0.484,	1.456)
	Type × Session	0	0.5	0	(-0.98,	0.98)
	Length × Session	0.136	0.516	0.264	(-0.876,	1.149)
	Type × Length × Session	-0.171	0.701	-0.244	(-1.545,	1.203)
L2 Group	Type	0.742	0.234	3.177**	(0.284,	1.2)
	Length	-0.017	0.217	-0.08	(-0.443,	0.408)
	Session	0.044	0.21	0.21	(-0.368,	0.457)
	Lg10frequency	1.09	0.137	7.959***	(0.821,	1.358)
	Type × Length	0.608	0.345	1.76	(-0.069,	1.285)
	Type × Session	-0.621	0.31	-2.004*	(-1.229,	-0.013)
	Length × Session	-0.173	0.285	-0.606	(-0.73,	0.385)
	Type × Length × Session	-0.16	0.451	-0.356	(-1.044,	0.723)

For the L1 group, the analysis did not return main effect for Type (p=.333), Length (p=.489), Session (p=1), or Lg10(frequency) (p=.071). Neither of the interactions of Type × Length (p=.326), Type × Session (p=1), Length × Session (p=.792), or Type × Length × Session

(p=.807) is significant. Pairwise contrast analyses also did not return any significant difference in NSs' judgments between idioms and FSs, 3- and 4-character stimuli, or two sessions.

For the L2 group, the analysis yielded main effect for Type (p=.002), Lg10(frequency) (p<.000), and the interaction of Type × Session (p=.045). No main effect was found for Length (p=.936), Session (p=.833), or the interaction of Type × Length (p=.078), Length × Session (p=.544), or the three-way interaction of Type × Length × Session (p=.722). Based on the significant results found for Type and Type × Session, pairwise contrasts were conducted for the L2 group with Type and Session as contrast fields. Regarding the stimuli type, the analysis returned significant difference for 3-FSs and 3-idioms in the Silent session (p=.011), 3-FSs and 3-idioms in the TA session (p<.000), and 4-FSs and 4-idioms in the TA session (p=.005), all with the FSs being judged more accurately than their idiom counterparts. Significance was not reached for 4-FSs and 4-idioms in the Silent session (p=.571). Concerning the contrasts by Session, NNSs judged 3-FSs (p=.003) and 4-FSs (p=.019) significantly different in the two sessions, both with the TA session being more accurate than the silent session. Judgments in the two sessions were not significantly different for 3-idioms (p=.505) or 4-idioms (p=.833).

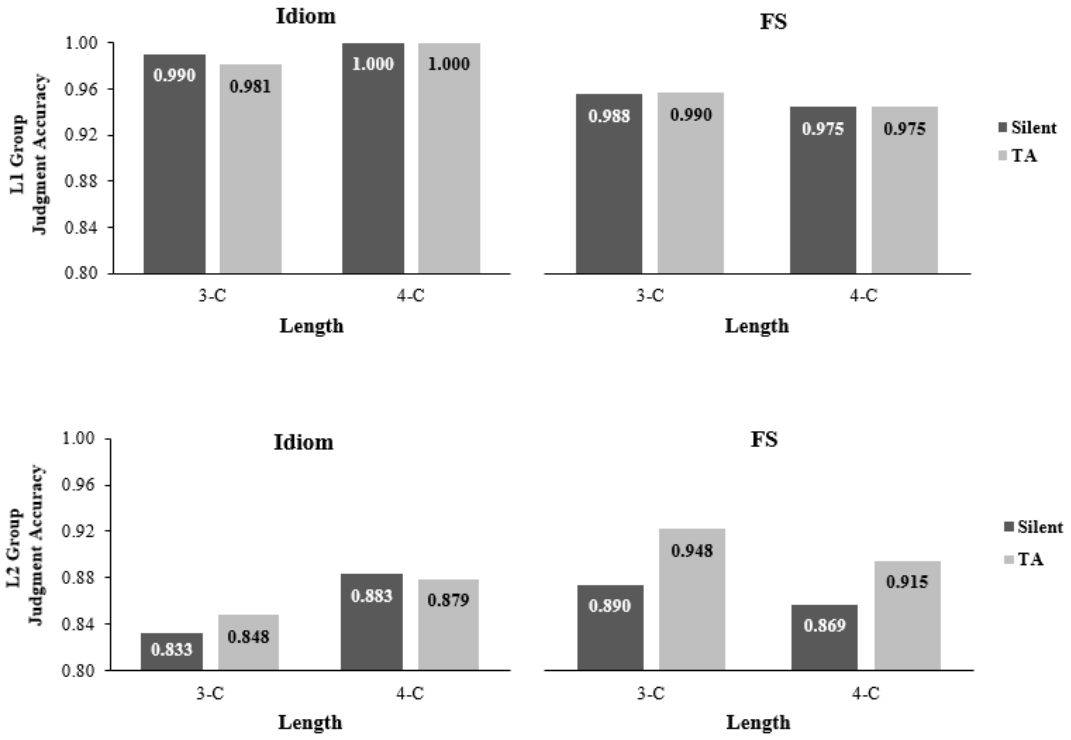

Figure 5.3: Judgment accuracy by group, type, length, and session

Figure 5.3 presents judgment accuracy on different stimuli in different sessions. Four processing patterns can be generalized: for the L1-silent session, 4-Idiom ≥ 3-Idiom ≥ 3-FS ≥ 4-FS (≥ indicating nonsignificantly more accurate than); for the L1-TA session: 4-Idiom ≥ 3-FS ≥ 3-Idiom ≥ 4-FS; for the L2-silent session, 3-FS ≥ 4-Idiom ≥ 4-FS ≥ 3-Idiom; for the L2-TA session, 3-FS ≥ 4-FS ≥ 4-Idiom ≥ 3-Idiom.

The second part of RQ2 asked ***do NSs and NNSs judge the same stimuli consistently in the two GJTs***? To respond to this question, the Yes/No judgments made for each item in the two sessions by each subject were compared. The 20 subjects in each group made a total of 960 judgments for each stimuli type in each session (48 items × 20 subjects=960). Table 5.5 presents the counts and ratios of the inconsistent judgments by group and stimuli.

Table 5.5: Judgment inconsistency between the two sessions by group and stimuli type

	L1 Group		L2 Group	
	Count	Ratio (out of 960)	Count	Ratio (out of 960)
3-Idiom	7	1.46%	49	10.21%
3-FS	7	1.46%	46	9.58%
4-Idiom	0	0	62	11.91%
4-FS	16	3.33%	46	9.58%

The results show that L1 speakers had 1.56% of the items judged inconsistently across the two sessions, and the inconsistency rate for the L2 group is 10.57%. The most significant discrepancy (11.91%) occurred in L2 learners judging 4-idioms. On the contrary, a level of 100% consistency was reached by L1 speakers judging 4-idioms, in which all 4-idioms were judged correctly in both sessions. In general, L1 speakers demonstrated a higher level of judgment consistency L2 learners.

RQ3. What do TA verbalizations reveal about L1 and L2 speakers' knowledge of idioms and FSs?

The first part of RQ3 asked ***can dichotomous judgments reflect NSs and NNSs' actual knowledge of idioms and FSs***. To answer this question, items that were provided with a correct Yes-or-No response in the TA session were extracted. TA verbalizations on these items were analyzed based on the four-way coding for the status of knowledge: "no evidence", "partial", "incorrect", and "correct". Incorrectly responded items in the TA session were not analyzed since an incorrect Yes-or-No response was by itself indicating "incorrect" knowledge. From Figure 5.4, two different patterns can be seen.

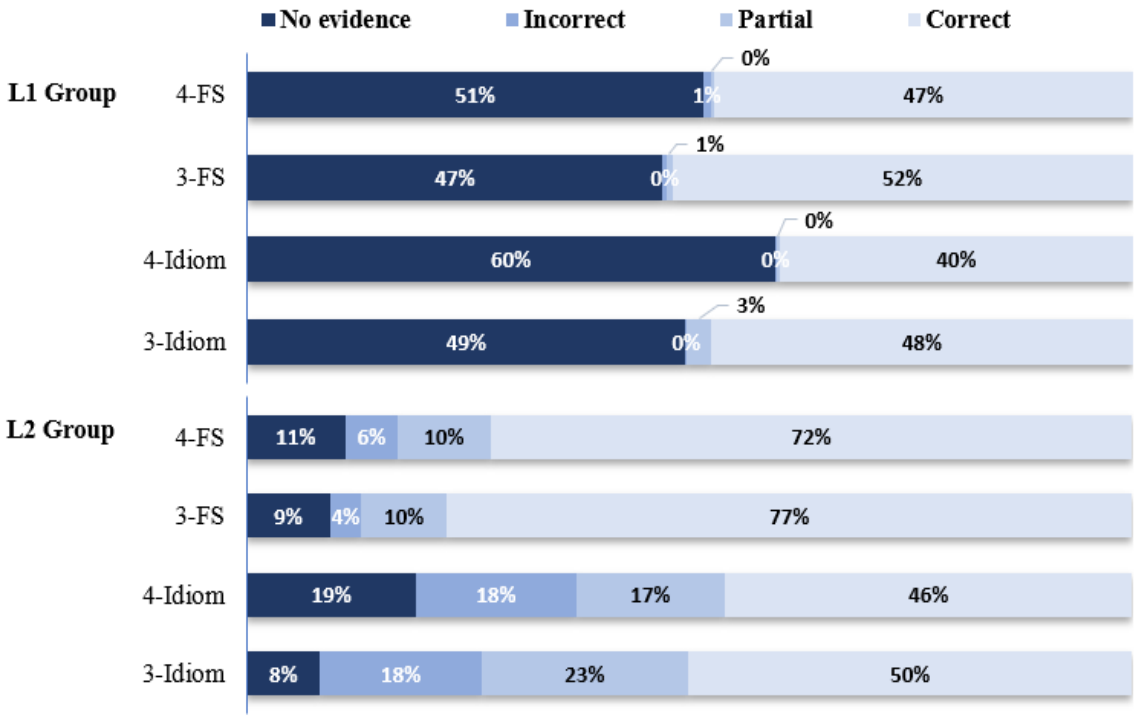

Figure 5.4: Distribution of the status of knowledge on correctly judged items by group and stimuli type

For the L1 group, most verbalizations were coded either as "no evidence" or as "correct"; a merely 1% to 3% TAs showed "partial" or "incorrect" knowledge. For the L2 group, "correct" knowledge accounted for 56% of the verbalizations on average; "partial" and "incorrect" knowledge accounted for 16% of the items on average, and another 12% TAs showed "no evidence" of the status of knowledge. Across all four types of stimuli, FSs were thought aloud more accurately by L2 learners than idioms.

Because only "incorrect" and "partial" responses can show tangible evidence that an item was misinterpreted, generalizations were made based on the distributions of the "partial" and "incorrect" TAs. For NSs, dichotomous judgments can reflect NSs' actual knowledge given the small proportion of "partial" or "incorrect" TAs. For the L2 group, the results of dichotomous judgments tended to overestimate learners' knowledge on all stimuli types given that 16% of the

stimuli were thought aloud partially correct or incorrectly. L2 learners' knowledge of idioms was notably limited given that 41% of the correctly judged 3-idioms and 35% of the correctly judged 4-idioms showed "partial" and "incorrect" knowledge. These findings suggest that the dichotomous judgment results didn't reflect L2 learners' actual knowledge of Chinese idioms.

To answer the second part of RQ3: ***do NSs and NNSs use the same strategies to process idioms and FSs***, a GLMM was computed with the coding of Strategy (holistic vs. analytical) as the target, Group, Type, and Length as fixed factors, and Subject and Item as random factors. Table 5.6 presents the results.

Table 5.6: Fixed coefficients for processing strategy

	Coefficient	SE	*t*	CI	
Group	-2.340	0.374	-6.258***	-3.074	-1.607
Length	-0.719	0.137	-5.248***	-0.988	-0.451
Type	-0.644	0.135	-4.772***	-0.908	-0.379
Group × Length	-0.423	0.285	1.487	-0.982	0.135
Group × Type	-0.215	0.268	-0.802	-0.741	0.311
Length × Type	0.633	0.193	3.436**	0.285	1.041
Group × Length × Type	0.236	0.415	0.57	-0.576	1.049

The main effect was found for Group (p<.000), Type (p<.000), and Length (p<.000). Significance was also obtained for the interaction of Type × Length (p=.001) but not for the interaction of Group × Type (p=.423) or Group × Length (p=.137). Pairwise contrasts showed that NSs used significantly more holistic strategies on 3-idioms than 4-idioms (p=.001). However, NNSs used similar amount of holistic and analytical strategies on 3- and 4-FSs (p=.38). The same patterns were observed in the L2 group, in which significantly more (17%) holistic strategies were used for 3-idioms than for 4-idioms (p<.000), while a similar amount of holistic and analytical strategies were used for 3- and 4-FSs (p=.683). A significant difference was observed between 4-FSs and 4-idioms for both the L1 (p=.004) and L2 group (p<.000), but

the difference between 3-FSs and 3-idioms was not significant for either the L1 (p=.887) or L2 group (p=.89). Figure 5.5 presents the distributions of processing strategies by group and stimuli type.

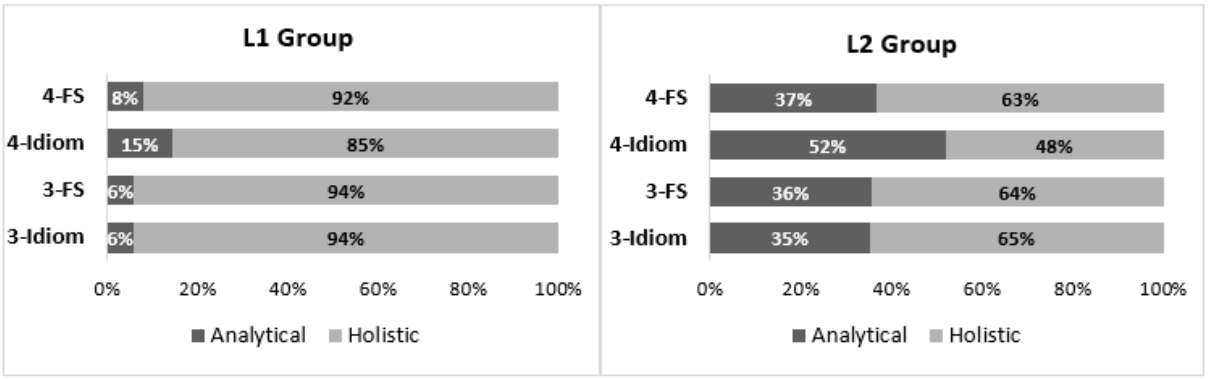

Figure 5.5: Distributions of processing strategies by group and stimuli type

From Figure 5.5, three processing strategy patterns can be observed. First, both groups processed the majority of stimuli holistically regardless of type or length. Second, both groups showed a notably different strategy pattern for 4-idioms from other types of stimuli with fewer holistic and more analytical strategies being used. Third, L1 speakers used significantly more holistic strategies and fewer analytical strategies than L2 learners did.

The last part of RQ3 asked: ***is processing strategy correlated with the accuracy of dichotomous judgments***? Spearman's correlations were performed for Strategy (scoring: 1 for holistic strategy and 0 for analytical strategy) and the judgments provided in the TA session (scoring: 1 for correct judgment and 0 for incorrect judgment). The L1 group showed a positive correlation between Strategy and Judgment with marginal significance (p=.048), suggesting that NSs' correct judgments were likely to be made through holistic processing than through analytical processing. The L2 group had a significant positive correlation (p=.01), indicating that L2 learners' correct judgments were also more likely to be made through holistic processing of the target forms than through analytical processing. In summary, the holistic strategy seemed to

facilitate both NSs and NNSs to make more correct judgments, while analyzing the internal elements increased the likelihood of making judgment errors.

Discussion

What Can Response Time Measurement Tell Us?

To summarize the findings of RT data, NSs responded to idioms and non-idiomatic FSs significantly faster than NNSs did. This result replicates the majority of findings in the literature (Jiang & Nekrasova, 2007; Siyanova-Chanturia et al., 2011; Schmitt et al., 2004; Schmitt & Underwood, 2004). Concerning the processing of LBs with different types and lengths, NSs and NNSs showed different patterns. NSs were more sensitive to type than length, processing idioms faster than FSs regardless of if the idioms contain 3 or 4 characters. NNSs were more sensitive to length than type, with 3-character LBs being significantly faster than 4-character ones regardless of if they were idioms or FSs.

The effect of type is an indicator of whether or not idioms and non-idiomatic FSs have the same lexical representations. The presence of a type effect in the L1 group suggests that idioms and non-idiomatic FSs are represented differently in NSs' mental lexicons. Although both are considered formulaic language, idioms are more like to be recognized as whole chunks and their meanings are directly retrieved, while FSs may not be processed in the same way as idioms are. The absence of a type effect in the L2 group indicates that idioms and FSs are not categorically different in NNSs' cognition. The different patterns found for NSs and NNSs is in support for Abel (2003) and Myles and Cordier's (2017) claim that simply because an LB is proved to be stored in NSs' lexicons holistically does not mean it has the same mental representation in NNSs' lexicons. In this study, the overall frequency of the idioms is higher than

the overall frequency of FSs. Therefore, NSs may encounter the idioms more often the FSs, which caused the fast recognition. However, this frequency difference did not make NNSs recognize the idioms more rapidly than the FSs, suggesting that frequency is not a determining factor for NNSs' idiom processing.

The effect of length is indicative of in what manner a sequence is recognized. If a sequence is read in a word-by-word fashion as the sequence unfolds from left to right, then length should show a main effect since 4-character sequences are one character longer than 3-character ones. The presence of length effect in the L2 group suggested that L2 learners read the idioms and FSs both in a verbatim fashion. The absence of length effect in the L1 group suggests that NSs did not read the LBs in a word-by-word manner. Cacciari and Tabossi (1988) proposed that the comprehension of LBs is an on-going process of configurating a string of words. When sufficient input allows a phrase to be recognizable as an idiom, the configuration halted and speakers retrieve the meaning associated with the idiom. If a phrase is not recognized as an idiom, the configuration continues until speakers derive a meaning from analyzing the individual words. For visually presented LBs, NSs' perceptual span are able to capture both 3- and 4-charater phrases by on gaze (Inhoff & Liu, 1997). Therefore, NSs can quickly recognize a phrase to be an idiom or not. If it is an idiom, NSs make a judgment promptly with configuring all the words; if it is an FS, NSs configurate the constituent words until the phrase makes sense to them. These two processes finally cause idioms to be processed faster than FSs by NSs. Because the FSs also enjoy some degree of formulaicity, we suspect that the configuration of FSs stop at some middle point when the input is sufficient for NSs to predict the rest part of an FS, resulting in no RT different found between 3- and 4-character FSs.

What can the dichotomous judgment measurement tell us?

Dichotomous judgment data were collected in both GJT sessions. NSs demonstrated a higher degree of judgment accuracy and consistency than NNSs in the two GJTs. The difference found between NSs and NNSs is consistent with that observed by Jiang and Nekrasova (2007). With regard to judgment accuracy between idioms and FSs, significance was not reached for NSs, indicating that both types of LBs were equally familiar to NSs. NNS, however, judged FSs more accurately and consistently than their idiom counterparts in two GJTs. The finding that NNSs understand FSs better than idioms is not completely unpredictable. First, the selected FSs are all commonly used in daily communicative situations. NNSs who learn Chinese in a target language environment more often hear and use the FSs than they hear and use the idioms. Although FSs and idioms have similar corpus frequency, idioms' corpus frequencies are mainly contributed by formal written sources such as newspaper articles or literary work that NNSs have limited exposure to. The usage-based model predicts that maximal and interactive exposure can enhance learners' sensitivity to a language (e.g., Abbot-Smith & Tomasello, 2006; Bannard & Lieven, 2009). The finding about NNSs lends evidence to this argument. Second, FSs are non-idiomatic novel phrases. Even if NNSs encountered an unfamiliar FS, they were able to understand the meaning of it through computing and integrating its constituent elements (Libben, 1998; Sandra, 1994) and make a correct judgment. However, computing the constituent words may not help NNSs to understand an unfamiliar idiom because idiom's meaning is not always derivable from its constituent words. Therefore, if learners depend on a verbatim interpretation to make sense of idioms (which was the case in this study as shown by the TA data), judgment errors will occur.

Concerning the reactivity issue for judgment accuracy, the nonsignificant *p*-values for session in both groups indicated that the TA procedure showed no significant reactivity effect on the NSs and NNSs' performance of the GJT task. This result echoes the findings of the meta-analysis performed by Bowles (2010), proving that the TA procedure is able to provide valid results that can be compared with other concurrent data. However, the pairwise contrasts revealed that NNS's judgment accuracy on FSs was higher than in the TA session than in the silent session. We suspect that this difference is due to that in the TA session, participants were asked to read aloud the stimuli first. Reading aloud the stimuli successfully drew NNSs' attentions to the forms. Therefore, some careless mistakes that NNSs might have made in the silent GJT session were avoided.

What can the think-aloud measurement tell us?

Status of knowledge

Based on how much knowledge speakers exhibited to have about LBs, TA verbalizations were coded into four categories: no evidence, incorrect knowledge, partial knowledge and correct knowledge. NSs' verbalizations were coded "correct", and the other half fell into the no-evidence category. No evidence cannot be equated with no knowledge. Most no-evidence TAs were NSs' intuitive comments, such as "this one is correct because we often use it", or "this is an idiom, so is correct". Intuitive comments are considered as a shallow processing that involves little amount of cognitive effort (Leow & Mercer, 2015). This finding again shows that the stimuli LBs are considered "cliché" to NSs. The results of NSs' TA data generally overlapped with the results of dichotomous judgment data, indicating that the dichotomous judgment measurement is able to reflect NSs' status of knowledge of the LBs.

In contrast, the results of NNSs' TA verbalizations showed some inconsistency with the dichotomous judgments. In dichotomous judgments, the accuracy rate of FSs was over 90%. TA data showed only 70% of FSs were thought aloud correctly, and 15% showed partial or incorrect knowledge. For idioms, the dichotomous judgment accuracy was above 85% while only 50% of the verbalizations on idioms showed fully correct knowledge, and another 43% showed partial or incorrect knowledge. These results demonstrated that the Yes/No judgment measurement tends to overestimate L2 learners' knowledge of both idioms and FSs. One cause of the overestimation could be the substantial amount of guesswork in the two-way judgement task (Birdsong, 1989; Mandell, 1999; N. Schmitt, D. Schmitt, & Clapham, 2001). Moreover, the inconsistency between the TA and dichotomous judgment measurements may also be caused by the different nature of the two tasks. The silent GJT which elicited dichotomous judgments is a recognition task and the TA procedure is essentially a production task. N. Ellis (2012) stated that the recognition of an LB is easier than the production of it, which could impact the outcomes of the studies. The no evidence TAs of NNSs provide evidence to this idea as NNS often admitted to "have learned" or "heard of" an LB but "forgot its meaning". Therefore, NNSs recognized an LB as a correct form in the silent GJT but failed to provide its meaning in the TA session. Just as Leow (1993) pointed out, the intake that learners use to perform immediate recognition "does not necessarily imply language acquisition" (p.334). Thus, the dichotomous judgments may not be sufficient to reflect the NNSs' actual status of L2 knowledge.

Through scrutinizing NNSs' TA data, three types of errors were observed. First, NNSs tended to overcorrect the forms by applying strict grammar rules. Consider Example 5.1.

Example 5.1

Target:	重要手段
(FS)	*zhongyao-shouduan* important-means "important means"
Subject 2:	这个不对…应该是'重要的手段'.
(Japanese L1)	"This one is incorrect…should be *zhongyao-de-shouduan.*"

The FS in Example 5.1 is a correct phrase. However, Subject 2 judged it to be incorrect and offered another correct phrase by applying the rule of adding a modification marker de between an adjective and a noun, where the de is often omittable. The overcorrection behavior may have to do with the type of input that NNSs had received. Although the learners are currently studying Chinese in China, their learning is still largely confined to classroom settings where textbooks and teacher talk are the major sources of input (Meunier, 2012). Textbooks present recursive rules, and teachers provide corrective feedback primarily to violations of these rules (Krashen 1976; Krashen & Seliger 1975; Pica, 1983; Terrell, 1991). This type of input often directs learners attention away from meaning and draw attentions to form and accuracy (Gregersen, 2003). Consequently, learners become habituated to inspecting the well-formedness of a new expression. However, in real-world communication, language forms can be flexible. Lacking of the real-world input and thereby a confident language intuition, NNSs relied on rigorous rules to make judgments which led to the overcorrection errors.

The second type of errors involves semantic inference, manifested as L2 learners over-extending the semantics of an LB. Consider the cases in Example 5.2.

Example 5.2

a. Target:	心里有事
(FS)	*xin-li-you-shi* heart-inside-have-thing "have something in one's mind."
Subject 6: (Korean L1)	没有听过, 但是好像就是有喜事的意思…我猜… 应该是对的. "Never heard of this, but seems like it just means 'have some blessed thing'…I guess…should be correct."
b. Target:	脱口而出
(FS)	*tuo-kou-er-chu* blurt-mouth-and-out "blurt out"
Subject 7: (Mongolian L1)	对的. 就是把秘密, 不知道的事情, 说着说着就说出来了. "Correct. Just means a secret, something people don't know, was spilt out while talking."

In Example 5.2a, the FS simply means "have something in one's mind". Based on Subject 6's verbalization, she was able to infer the meaning of the expression. However, instead of adhering to the literal meaning, she extended the surface meaning to "have some blessed thing", where the sense of "blessed" does not exist. The semantic extension has also been observed for idioms as demonstrated in Example 5.2b, where the subject extended the idiom's meaning from "blur out" to "blurt out a secret". Research investigating learners' lexical inference when dealing with unknown texts (Nassaji, 2003; de Bot, Paribakht, & Wesche, 1997; Frantzen, 2003; Morrison, 1996) has derived some explanations for how lexical inference has been done. Learners' pre-existing knowledge and world knowledge were found to play important parts (Nagy, 1997). We argue that learners' pre-linguistic knowledge may trigger semantic overextensions: learners may have read or heard similar expressions in particular contexts and used that experience to infer the meaning of a new expression. That the overextension phenomenon was often accompanied by the description of a scenario may serve as evidence in support of this assumption. Besides the prior knowledge on L2, learners' L1 knowledge could also impact the understanding of LBs (Yamashita & Jiang, 2010; Iruho, 1986; Cooper;1999).

Unfortunately, because participants were from different L1 backgrounds, it was difficult to identify the influence of a negative L1-to-L2 transfer without speakers mentioning it explicitly. Based on the observation of positive L1-to-L2 transfer evidence (see Example 5.3), we can only speculate that negative L1 to L2 transfer that caused the semantic inference errors.

Example 5.3

Target:	千方百计
(Idiom)	*qian-fang-bei-ji*
	thousand-method-hundred-strategy
	"(to take) every means to"
Subject 19:	对的.韩语也有这个, 就是用很多方法做.
(Korean L1)	"Correct. Korean also has this saying, just meaning 'use many means to do'".

Another possible cause of lexical inference errors may be related to the stimuli list context (Ferrand & Grainger, 1996; Klauer, Roßnagel, & Musch, 1997). If participants perceived that a considerable number of stimuli are idioms, they are inclined to interpret the unfamiliar stimuli idiomatically. Since NSs' performance did seem to be affected by the stimuli list context, we argue that being affected by the experiment context is also a manifestation of the instable status of L2 knowledge.

The third type of semantic-based errors is related to processing strategies used by NNSs, which will be discussed in the next section.

Processing strategies

TA data were coded into holistic and analytical based on the two hypotheses of how LBs are processed. Since the process of comprehension is not directly assessible, speakers' concurrent verbalizations may provide a window to infer the process. The results revealed NSs used holistic strategies for over 90% of the LBs while NNSs only used holistic strategies for approximately 60% of the LBs. This pattern is in conformity with that observed by Abel (2003), that NNSs tended to think an LB to be decomposable. Schweigert (1986) found that speakers did

analysis to idioms that are not familiar to them. Therefore, The low familiarity could be the major reason of why NNSs used more analytical strategies than NSs. Because of the low familiarity, an idiom may be harder to retrieve even though it has been previously encountered. When the direct retrieve failed, NNSs had to rely on analyzing the internal components to infer the meaning of the idiom. Just as Abel (2003) reported, in an exit survey after an idiom comprehension task, German L1 English L2 learners said if they encountered an unknown idiom, they tried to put together the literal meanings of the constituent words to derive a meaning. The present study replicated Abel's finding and further found that the act of analysis could cause comprehension errors. Consider Example 5.4.

Example 5.4

a. Target (Idiom):	谈天说地 *tan-tian-shuo-di* Talk-sky-speak-earth "talk of everything under the sun"
Subject 11: (Thai L1)	对的. 就是谈谈天气. "Correct. Just means 'talk about the weather.'"
b. Target (Idiom):	出人命 *chu-ren-ming* happen-human-life "death-causing (accident)"
Subject 9: (Japanese L1)	对的. 意思就是出生了一个人. "Correct. The meaning is 'give birth to a person.'"

As it can be seen in both cases in Example 5.4, subjects literally interpreted the constituent words and derived an incorrect interpretation for the idioms. The statistics showed that analytical strategy was significantly correlated with NNSs' judgment error ($p=.01$). This finding first suggested although the idioms were rated by CSL teachers to be eligible for HSK6 learners, the idiom items are equally familiar to all the L2 participants. This finding also lent

evidence to the claim that to understand an idiom, speakers must have learned it (Bobrow & Bell, 1973; Swinney & Cutler, 1979; Gibbs, 1980; Cacciari & Glucksber, 1991).

Regarding the processing strategy associated to different types of LBs, more analytical strategies were used for 4-idioms than 3-idioms by both NSs and NNSs. This distribution may be impacted the decomposability nature of the idioms. A post hoc analysis was conducted based on the descriptive norms collected in Study I. We found that the selected 4-idioms were rated more decomposable than the 3-idioms by NSs ($t=-4.578$, $p<.000$). Appendix F displays the average ratings on the compositionality of the selected 3- and 4-idioms. Gibbs, Nayak, and Cutting (1989) proposed that speakers have intuitions about how the meanings of the individual words contribute to the figurative meaning of an idiom. We argue that speakers' processing strategies just reflected such intuitions that more decomposable idioms were analytically verbalized and less decomposable idioms were holistically verbalized. Another finding that seems contradictory to the decomposability argument is that both NSs and NNSs used more holistic strategies in 4-FSs than 4-idioms. Since FSs are fully decomposable phrases, they are supposed to be processed analytically. By a closer scrutiny, we found that NSs verbalizations on 4-FSs were mainly coded "no evidence", suggesting that NSs considered 4-FSs to be everyday cliché; NNSs often provided an example sentence for an 4-FS; indicating NNSs may use the 4-FSs frequently. These findings demonstrate that some compositional configurations, though not being idiomatic, still become entrenched in language users' memories due to the high frequency and familiarity (Siyanova-Chanturia, 2015; N. Ellis, 2012).

General Discussion

Comparing the results for the RT data and the TA data, different processing patterns for the 4-idioms were observed in the L1 group. The RT data showed that 4-idioms were processed significantly faster than any other types of sequences by L1 learners, which is usually considered evidence of holistic processing (e.g., Jiang & Nekrasova, 2007). However, the TA data seem to suggest that 4-idioms were processed analytically. We suggest that these results are not contradictory; instead, the difference is caused by different measurements. In the silent GJT, as long as the participants could visually recognize that the target form was something familiar to them, they could press the button to indicate their judgment. Due to this task procedure, NSs' processing may have stopped at successful recognition of the written patterns without progressing to the semantic processing stage (Laberge & Samuels, 1974). In other words, in the silent GJT, the visual stimuli had not yet transformed into meanings. Pattern recognition often occurs in automatic processing that requires readers to be highly fluent, and the stimuli are of high saliency. Idioms are an obvious type of lexical bundle and are noticeably identifiable as single units (Schmitt & Carter, 2004). The 4-idioms are the prototype for Chinese idioms, as 98.75% of Chinese idioms have four characters. Therefore, the 4-idioms could be quickly recognized by NSs in silent GJT without having to invoke their meanings. However, in the TA GJT, NSs moved from the visual processing stage to the semantic processing stage, and it was at this level that the factors, such as register and compositionality, came into play. In other words, the RT data and the TA data tapped into different processing stages for NSs. In contrast, L2 learners had not reached high automaticity in reading and thus might not have been able to depend on pattern recognition to make all their judgments. Therefore, semantic processing was evoked in both GJT sessions for them. As a result, the two types of data showed similar patterns

regarding how L2 learners process different types of stimuli. These findings again demonstrate that RT data and TA data are complementary and reveal a more comprehensive picture of idiom processing.

CHAPTER 6: CONCLUSION AND IMPLICATIONS

This dissertation investigates the processing of two types of Chinese idioms by native speakers and second language learners. The general questions addressed in this dissertation are:

1. What are speakers' intuitions about the lexical properties of the two types of idioms?

2. Will the literal meaning of the constituent word be activated during the processing of the two types of idioms?

3. What strategies (holistic or analytical) do L1 and L2 speakers use to process idioms and non-idiomatic phrases?

Three studies conducted in this dissertation employ three measurements-metalinguistic judgments/ratings, response times, and think-aloud protocols-to answer these questions.

Metalinguistic ratings are collected in Study 1. The study contributes a database of native speakers' ratings for the 425 most commonly used idioms in six linguistic dimensions: familiarity (how often an idiom is encountered), meaningfulness (how well an idiom is understood), compositionality (how much do constituent words contribute to the overall meaning of an idiom), literality (how likely is the literal interpretation of an idiom is plausible in real world), final-word predictability (how likely is an idiom fragment with the last word missing is completed as the expected idiom), and linguistic register (whether an idiom is used more often in the formal written context or in the informal spoken context). The average ratings show that GYYs are scored lower than CYs in every dimension, providing psycholinguistic accounts for the categorization of GYYs and CYs as distinct types of idioms.

Metalinguistic judgments are also collected in a GJT in Study 3, where L1 and L2 speakers are asked to make yes-or-no decisions on whether the phrases (idioms or non-idiomatic

FSs) are grammatically acceptable Chinese. The results show that L1 speakers are equally accurate in judging idioms and their matched non-idiomatic FSs. However, descriptive data reveal that idioms are judged slightly more accurately than FSs, and CYs are judged more accurately than GYYs are. These findings suggest that CYs, GYYs, and FSs, may lie closely along a lexicalization gradient where one type is slightly more lexicalized than the other type. For L2 speakers, FSs were judged more accurately than idioms, and CYs were judged more accurately than GYYs. These findings perfectly align with the idiom decomposition hypothesis (IDH), which assumes that more compositional phrases have the processing advantage over the less compositional ones. In this study, the select non-idiom FSs are more compositional than idioms, and the CYs are more compositional than GYYs. The reason is that because the verbatim analysis of a compositional phrase matches its overall meaning, the analysis of the phrase will be more quickly verified. In other words, IDH assumes that people process idioms the same way as they process novel phrases. The IDH's accounting for L2 learners' processing rather than for L1 speakers' processing may indicate that compositionality is not the essential determiner of the L1 idiom processing.

RT data are collected in the same GJT session in Study 3. The results show that L1 speakers process idioms significantly faster than FSs; CYYs are processed insignificantly faster than GYs. L2 speakers, on the other hand, process phrases of the same length equally fast regardless of whether they are idioms or FSs. The findings are indicative of how L1 and L2 lexicons are different from each other, and the two populations may use different models or strategies to process the multi-word lexical bundles.

Another set of RT data are collected in Study 2, where native speakers make lexical decisions on a disyllabic word (DSW) primed by an idiom whose second constituent word is

semantically related to the DSW or a novel phrase containing the same second constituent word with the related idiom or an idiom that does contain any related constituent words with the DSW. The rationale is that if the second constituent word is activated, the related constituent words will provide semantic cues for the target DSWs, and thus, the RTs to the DSWs when primed by the related idioms and related novel phrases will be shorter than when the DSWs are primed by unrelated idioms. The results demonstrate significant priming effects for GYYs but not for CYs, indicating that the literal meaning of the second constituent words in GYYs is activated whereas the second constituent words in CYs are not activated. The results are consistent with those revealed by the RT data collected in the GJT session in Study 3, where GYYs are processed slower (though insignificantly) than CYs. These findings may indicate that the lexical representations of GYYs and CYs may be different, with GYYs aligned with novel phrases and CYs inclined to frozen words. The difference between GYYs and CYs found by the RT data also accords with the lexical and syntactic properties of the two idiom types; GYYs are more syntactically flexible and subject to lexical substitution, while CYs are completely frozen forms, rejecting any internal alternation.

Finally, L1 and L2 speakers' verbalizations are collected in the TA-GJT session, where participants think aloud, speaking their minds when they make yes-or-no judgments on idioms or non-idiomatic FSs. L1 speakers' TAs show the same processing pattern shown by the silent GJT session, while L2 speakers' verbalizations show incomplete or incorrect knowledge of the idioms that they judged accurately in the silent GJT session. Moreover, L1 speakers use significantly more holistic strategies than L2 learners do in processing both types of phrases. This finding echoes that revealed by the RT data in the silent GJT session that L1 and L2 speakers use

different strategies to process multi-word lexical bundles. The TA data also confirm that RT data are more precise in revealing the processing patterns than dichotomous judgment data.

The diverse measures used in this dissertation make some novel contributions to the investigations of idiom processing and implications to the future research.

First, the studies provide detailed descriptions of lexical properties of the two types of idioms and offer reliable statistics for future studies on Chinese idioms.

Second, a working hypothesis can be proposed regarding the lexical status of GYYs and CYs that GYYs may be more like phrases while CYs are more like single words from the speaker-internal perspective. However, more comprehensive understanding of the lexical status of the two types of idioms can only be established on a series of empirical studies controlling more meticulous parameters. Future research could contrast idioms with matched novel phrases and matched single words. With the phrase-idiom-word three-way contrast, we can go one step further in discussing the lexical status and mental representations between phrases and idioms as well as between idioms and single words.

Third, although the English word "idiom" is translated into Chinese as "chengyu" ("fixed language"), which only refers to CY, another type of Chinese idioms, GYY was also found to be commonly used by native speakers. Because GYYs possess all the characteristics that are observed in idioms, this type of idioms also merits a closer investigation. Besides, L2 learners also demonstrated problems in processing the three-character idioms (GYYs), which calls for attention to the teaching of GYYs.

Fourth, introspective data elicited from the think-aloud procedure showed that L2 learners' knowledge of idioms or, in general, lexical bundles could be partial or incorrect, and this fact was reflected by quantitative data (e.g., RTs and Yes-or-No judgments) used to measure

L1 speakers' processing. The finding suggests that further LB processing with L2 participants should consider triangulating the data that can reflect the processing pattern and the depth of knowledge of L2 learners to have a more comprehensive understanding of SLA.

Finally, the introspective data also revealed some problems with idiom acquisition in L2 that are worthy of further investigation. In L2 speakers' think-aloud verbalizations, we found a certain amount of "meaning forgotten" idioms. These idioms were claimed to have been learned, but the meanings were forgotten. This phenomenon raises the issue of idiom learning and retention. Alali and Schmitt (2012) discussed whether teaching formulaic sequences should be the same as or different from teaching single words. The authors found that the learning gain of idioms was lower than the learning gain of single words. The question is why the retention of idioms is more difficult than the retention of single words. Future intervention studies could compare the effects of different approaches to teaching Chinese idioms.

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.

Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second Language Research*, *19*(4), 329-358.

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

Agha, A. (2004). Registers of language. In A. Duranti (Eds.) *A companion to linguistic anthropology*, 23-45. John Wiley and Sons.

Alali, F. A., & Schmitt, N. (2012). Teaching formulaic sequences: The same as or different from teaching single words?. *TESOL Journal*, *3*(2), 153-180.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67-82.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284-17289.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*(1), 137-172

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, 26, 181-190.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Birdsong, D. (1992). Ultimate attainment in second language acquisition. *Language*, 706-755.

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.

Bonin, P., Méot, A., & Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behavior Research Methods*, *45*(4), 1259-1271.

Bowles, M. A. (2004). L2 glossing: To CALL or not to CALL. *Hispania*, 541-552.

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.

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

Bowles, M. A. (2010b). Concurrent verbal reports in second language acquisition research. *Annual Review of Applied Linguistics*, *30*, 111-127.

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.

Bulkes, N. Z., & Tanner, D. (2017). "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods*, *49*(2), 772-783.

Burt, J. S. (1992). Against the lexical representation of idioms. *Canadian Journal of Psychology*, *46*(4), 582.

Bybee, J., & Hopper, P. (2001). Of linguistic structure. In J. Bybee and P. Hopper (Ed.) *Frequency and the emergence of linguistic structure*, 1-24. John Benjamins Publishing.

Cacciari, C., & Glucksberg, S. (1995). Understanding idioms: Do visual images reflect figurative meanings?. *European Journal of Cognitive Psychology*, *7*(3), 283-305.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of memory and language*, *27*(6), 668-683.

Cacciari, C., Padovani, R., & Corradini, P. (2007). Exploring the relationship between individuals' speed of processing and their comprehension of spoken idioms. *European Journal of Cognitive Psychology*, *19*(3), 417-445.

Caillies, S., & Butcher, K. (2007). Processing of idiomatic expressions: Evidence for a new hybrid view. *Metaphor and Symbol*, *22*(1), 79-108.

Canal, P., Pesciarelli, F., Vespignani, F., Molinaro, N., & Cacciari, C. (2017). Basic composition and enriched integration in idiom processing: An EEG study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(6), 928-943.

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.

Carrol, G., & Conklin, K. (2019). Is All Formulaic Language Created Equal? Unpacking the Processing Advantage for Different Types of Formulaic Sequences. *Language and speech*, 0023830918823230.

Chen, H. C., & Shu, H. (2001). Lexical activation during the recognition of Chinese characters: Evidence against early phonological activation. *Psychonomic Bulletin & Review*, *8*(3), 511-518.

Chiarello, C., Burgess, C., Richards, L., & Pollock, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't… sometimes, some places. *Brain and Language*, *38*(1), 75-104.

Chinese Academy of Social Sciences. (2015). The contemporary Chinese dictionary (5th edition) [J.] 现代汉语词典 (第五版). The Commercial Press, Beijing.

Chung, K. K. H., Code, C., & Ball, M. J. (2004). Lexical and non-lexical speech automatisms in aphasic Cantonese speakers. *Journal of Multilingual Communication Disorders*, 2(1), 32-42.

Cieślicka, A. B. (2015). Idiom acquisition and processing by second/foreign language learners. *Bilingual Figurative Language Processing*, 208-244.

Clackson, J. (2010). Colloquial language in linguistic studies. In Eleanor Dickey, Anna Chahoud (Eds.) Colloquial and Literary Latin, 7-11. Cambridge University Press.

Cohen, A. D. (2000). Exploring strategies in test taking: Fine-tuning verbal reports from respondents. In G. V. Ekbatani and H. D. Pierson (Ed.) *Learner-directed assessment in ESL*, 127-150. Routledge.

Conklin, K., & Carrol, G. (2018). First Language Influence on the Processing of Formulaic Language in a Second Language. In A. Siyanova-Chanturia and A. Pellicer-Sanchez (Eds.) *Understanding formulaic language: A second language acquisition perspective. Routledge.*

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1084-1096.

Conrad, S. M., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*, 56-70.

Cooper, T. C. (1999). Processing of idioms by L2 learners of English. *TESOL quarterly*, *33*(2), 233-262.

Corp, I. B. M. (2017). IBM SPSS Statistics: Version 25.

Cronk, B. C., & Schweigert, W. A. (1992). The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied Psycholinguistics*, *13*(2), 131-146.

Cronk, B. C., Lima, S. D., & Schweigert, W. A. (1993). Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research*, *22*(1), 59-82.

Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blendsß. *Memory and Cognition*, *25*(1), 57-71.

Davis, J. N., & Bistodeau, L. (1993). How do L1 and L2 reading differ? Evidence from thinking aloud protocols. *Modern Language Journal*, *77*(4), 459-472.

De Bot, K., Paribakht, T. S., & Wesche, M. B. (1997). Toward a lexical processing model for the study of second language vocabulary acquisition: Evidence from ESL reading. *Studies in Second Language Acquisition*, 309-329.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283-321.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual review of applied linguistics*, *32*, 17-44.

Ellis, N. C., Simpson-Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, *42*(3), 375-396.

Ellis, R. (1991). Grammatically judgments and second language acquisition. *Studies in Second Language Acquisition*, *13*(2), 161-186.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29-62.

Evett, L. J., & Humphreys, G. W. (1981). The use of abstract graphemic information in lexical access. *The Quarterly Journal of Experimental Psychology*, *33*(4), 325-350.

Fanari, R., Cacciari, C., & Tabossi, P. (2010). The role of idiom length and context in spoken idiom comprehension. *European Journal of Cognitive Psychology*, *22*(3), 321-334.

Ferguson, C. A. (1994). Dialect, register, and genre: Working assumptions about conventionalization. In D. Biber and E. Finegan (Eds.) *Sociolinguistic perspectives on register, 15-30.* Oxford University Press on Demand.

Ferrand, L., & Grainger, J. (1996). List context effects on masked phonological priming in the lexical decision task. *Psychonomic Bulletin & Review*, *3*(4), 515-519.

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative health research*, *3*(4), 430-441.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of experimental psychology: Learning, Memory, and Cognition, 10*(4), 680-698.

Frantzen, D. (2003). Factors affecting how second language Spanish students derive meaning from context. *The Modern Language Journal*, *87*(2), 168-199.

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

Fraser, B. (1999). What are discourse markers?. *Journal of Pragmatics, 31*(7), 931-952.

Fraser, C. A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition, 21*(2), 225-241.

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, *41*(2), 339-359.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive psychology*, *45*(2), 220-266.

Gass, S. M. (1994). The reliability of second-language grammaticality judgments. *Research Methodology in Second Language Acquisition*, 303-322.

Gass, S. M., & Selinker, L. (1983). *Language transfer in language learning: Issues in second language research*. Newbury House Publishers, Inc.

Gibbs Jr, R. W., & Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive psychology*, *21*(1), 100-138.

Gibbs Jr, R. W., & O'Brien, J. E. (1990). Idioms and mental imagery: The metaphorical motivation for idiomatic meaning. *Cognition*, *36*(1), 35-68.

Gibbs Jr, R. W., Nayak, N. P., & Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, *28*(5), 576-593.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & cognition*, *8*(2), 149-156.

Gibbs, R. W., Nayak, N. P., Bolton, J. L., & Keppel, M. E. (1989). Speakers' assumptions about the lexical flexibility of idioms. *Memory & Cognition*, *17*(1), 58-68.

Giora, R., & Fein, O. (1999). On understanding familiar and less-familiar figurative language. *Journal of Pragmatics, 31*(12), 1601-1618.

Glass, A. L. (1983). The comprehension of idioms. *Journal of Psycholinguistic Research*, *12*(4), 429-442.

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports and vocabulary learning scores. *Noticing and second language acquisition: Studies in honor of Richard Schmidt*, 183-205.

Gregersen, T. S. (2003). To err is human: A reminder to teachers of language-anxious students. *Foreign Language Annals*, *36*(1), 25-32.

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter?. *Language Learning*, *66*(2), 296-323.

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics*, *10*(2), 221-240.

Hamblin, J. L., & Gibbs, R. W. (1999). Why you can't kick the bucket as you slowly die: Verbs in idiom comprehension. *Journal of Psycholinguistic research*, *28*(1), 25-39.

Hoosain, R. (1992). Psychological reality of the word in Chinese. *Advances in Psychology*, *90*, 111-130.

Inhoff, A. W., & Liu, W. (1997). The perceptual span during the reading of Chinese text. *Cognitive processing of Chinese and related Asian languages*, 243-266.

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.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*(3), 433-445.

Kamimoto, T. (2008). Nation's Vocabulary Levels Test and its successors: a re-appraisal. Unpublished PhD Thesis. University of Wales: Swansea.

Kecskes, I. (2015). Is the idiom principle blocked in bilingual L2 production. *Bilingual figurative language processing*, *28*, 53.

Keysar, B., & Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans?. *Journal of Memory and Language*, *34*(1), 89-109.

Keysar, B., & Bly, B. M. (1999). Swimming against the current: Do idioms reflect conceptual structure?. *Journal of Pragmatics*, *31*(12), 1559-1578.

Kim, J. E., & Nam, H. (2017). The pedagogical relevance of processing instruction in second language idiom acquisition. *International Review of Applied Linguistics in Language Teaching*, *55*(2), 93-132.

Klauer, K. C., Rossnagel, C., & Musch, J. (1997). List-context effects in evaluative priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 246.

Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*(1), 68-101.

Krashen, S. D. (1976). Formal and informal linguistic environments in language acquisition and language learning. *Tesol Quarterly*, 157-168.

Krashen, S. D., & Seliger, H. W. (1975). The essential contributions of formal instruction in adult second language learning. *Tesol Quarterly*, 173-183.

Kroll, J. F. (1993). Accessing conceptual representations for words in a second language. *The Bilingual Lexicon*, *53*, 481.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*(2), 293-323.

Lakoff, G. (1986). The meanings of literal. *Metaphor and Symbol*, *1*(4), 291-296.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Lawson, M. J., & Hogben, D. (1996). The vocabulary-learning strategies of foreign-language students. *Language Learning*, *46*(1), 101-135.

Lennon, P. (1998). Approaches to the teaching of idiomatic language. *IRAL-International Review of Applied Linguistics in Language Teaching*, *36*(1), 11-30.

Leow, R. P. (1993). To simplify or not to simplify: A look at intake. *Studies in Second Language Acquisition*, *15*(3), 333-355.

Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning*, *47*(3), 467-505.

Leow, R. P. (1998). Toward operationalizing the process of attention in SLA: Evidence for Tomlin and Villa's (1994) finegrained analysis of attention. *Applied Psycholinguistics*, *19*(1), 133-159.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in Second Language Acquisition*, *22*(4), 557-584.

Leow, R. P. (2001). Attention, awareness, and foreign language behavior. *Language Learning*, *51*, 113-155.

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.

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.

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.

Leow, R. P., Hsieh, H. C., & Moreno, N. (2008). Attention to form and meaning revisited. *Language Learning*, *58*(3), 665-695.

Li, M., Jiang, N., & Gor, K. (2017). L1 and L2 processing of compound words: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition*, *20*(2), 384-402.

Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, *61*(1), 30-44.

Lindstromberg, S., & Boers, F. (2008). The mnemonic effect of noticing alliteration in lexical chunks. *Applied Linguistics*, *29*(2), 200-222.

Liu, Y., Li, P., Shu, H., Zhang, Q., & Chen, L. (2010). Structure and meaning in Chinese: An ERP study of idioms. *Journal of Neurolinguistics*, *23*(6), 615-630.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback?. *Studies in Second Language Acquisition*, *22*(4), 471-497.

Mandell, P. B. (1999). On the reliability of grammaticality judgement tests in second language acquisition research. *Second Language Research*, *15*(1), 73-99.

Mateu, J., & Espinal, M. T. (2007). Argument structure and compositionality in idiomatic constructions. *The Linguistic Review*, *24*(1), 33-59.

McCaskey, M. (1994). Teaching Japanese idioms using hypercard: an idiom module for a learner's dictionary. *Computer Assisted Language Learning*, *7*(2), 99-106.

McGlone, M. S., Glucksberg, S., & Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse Processes*, *17*(2), 167-190.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 558.

Meunier, F. (2012). Formulaic language and language teaching. *Annual Review of Applied Linguistics*, *32*, 111-129.

Millar, N. (2010). The processing of malformed formulaic language. *Applied Linguistics*, *32*(2), 129-148.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. *Bartning, M. Martin and I. Vedder. (Eds). Communicative proficiency and linguistic development: Interactions between SLA and language testing research*, 211-232. Eurosla.

Moon, R. (1998). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Eds.) *Phraseology: Theory, Analysis and applications*, 79-100. Oxford: Clarendon Press.

Morgan-Short, K., Heil, J., Botero-Moriarty, A., & Ebert, S. (2012). Allocation of attention to second language form and meaning: Issues of think-alouds and depth of processing. *Studies in Second Language Acquisition*, *34*(4), 659-685.

Morrison, L. (1996). Talking about words: A study of French as a second language learners' lexical inferencing procedures. *Canadian Modern Language Review*, *53*(1), 41-75.

Mueller, R. A., & Gibbs, R. W. (1987). Processing idioms with multiple meanings. *Journal of Psycholinguistic Research*, *16*(1), 63-81.

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.

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreignlanguage learning. *Language Learning*, *48*(3), 323-364.

Nagy, W. (1997). On the role of context in firstand second-language vocabulary learning. In N. Schmidt & M. McCarthy (Eds.), Vocabulary: Description, acquisition and pedagogy (pp. 64–83). Cambridge: Cambridge University Press.

Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *Tesol Quarterly*, *37*(4), 645-670.

Nassaji, H. (2006). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Modern Language Journal*, *90*(3), 387-401.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

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

Nayak, N. P., & Gibbs, R. W. (1990). Conceptual knowledge in the interpretation of idioms. *Journal of Experimental Psychology: General*, *119*(3), 315.

Needham, W. P. (1992). Limits on literal processing during idiom interpretation. *Journal of Psycholinguistic Research*, *21*(1), 1-16.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648-654.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology: general*, *106*(3), 226.

Neely, J. H., & Keefe, D. E. (1989). Semantic context effects on visual word processing: A hybrid prospective-retrospective processing theory. *Psychology of Learning and Motivation*, 24, 207-248.

Nekrasova, T. M. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning*, *59*(3), 647-686.

Nippold, M. A., & Taylor, C. L. (2002). Judgments of idiom familiarity and transparency. *Journal of Speech, Language, and Hearing Research*.

Nordmann, E., Cleland, A. A., & Bull, R. (2014). Familiarity breeds dissent: Reliability analyses for British-English idioms on measures of familiarity, meaning, literality, and decomposability. *Acta Psychologica*, *149*, 87-95.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, *70*(3), 491-538.

Oppenheim, N. (2000). The importance of recurrent sequences for non-native speaker fluency and cognition. In H. Riggenbach (Ed.), *Perspectives on Fluency*, 220–240.

Ortony, A., Schallert, D. L., Reynolds, R. E., & Antos, S. J. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior*, *17*(4), 465-477.

Pavlenko, A. (1999). New approaches to concepts in bilingual memory. *Bilingualism: Language and Cognition*, *2*(3), 209-230.

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.

Peterson, R. R., Burgess, C., Dell, G. S., & Eberhard, K. M. (2001). Dissociation between syntactic and semantic processing during idiom comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1223.

Pica, T. (1983). Adult acquisition of English as a second language under different conditions of exposure. *Language learning*, *33*(4), 465-497.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912.

Popiel, S. J., & McRae, K. (1988). The figurative and literal senses of idioms, or all idioms are not used equally. *Journal of Psycholinguistic Research*, *17*(6), 475-487.

Potter, M. C., So, K. F., Von Eckardt, B., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 23-38.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282-308.

Ravid, D., & Berman, R. (2009). Developing linguistic register across text types: The case of Modern Hebrew. *Pragmatics & Cognition*, *17*(1), 108-145.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355-371.

Read, J., & Nation, P. (2004). Measurement of formulaic sequences. In N. Schmitt (Eds.) *Formulaic Sequences: Acquisition, Processing and Use, 23-36.*

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*(2), 299-334.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89-104.

Roelofs, A. (2003). Modeling the relation between the production and recognition of spoken word forms. In N. O. Schiller and A. Meyer (Eds.) *Phonetics and phonology in language comprehension and production: Differences and similarities*, 115-158. Walter de Gruyter.

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, *25*(5), 762-776.

Rosa, E. M., & Leow, R. P. (2004). Awareness, different learning conditions, and second language development. *Applied Psycholinguistics*, *25*(2), 269-292.

Rosa, E., & O'Neill, M. D. (1999). Explicitness, intake, and the issue of awareness: Another piece to the puzzle. *Studies in Second Language Acquisition*, *21*(4), 511-556.

Rott, S. (2009). The effect of awareness-raising on the use of formulaic constructions. In R. *Corrigan, E. A. Moravcsik, H. Ouali and K. Wheatley (Eds.) Formulaic language: Acquisition, loss, psychological reality, and functional explanations*, 405-418. John Benjamins Publishing.

Salton, G. (2017). *Representations of Idioms for natural language processing: Idiom type and token identification, language modelling and neural machine translation*. Doctoral thesis, DIT, 2017. doi.org/10.21427/D77H8K.

Sandra, D. (1994). The morphology of the mental lexicon: Internal word structure viewed from a psycholinguistic perspective. *Language and Cognitive Processes, 9*(3), 227-269.

Schmitt, N. (2012). Formulaic language and collocation. In C. Chapelle (Eds.) *The encyclopedia of applied linguistics,* 1-10. John Wiley and Sons, Inc.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action. In N. Schmitt (Eds.) *Formulaic sequences: Acquisition, processing and use,* 1-22. John Benjamins Publishing.

Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Eds.) *Formulaic sequences: Acquisition, processing and use*, 173-189. John Benjamins Publishing.

Schmitt, N., Dornyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences. In N. Schmitt (Eds.) *Formulaic sequences: Acquisition, processing and use,* 55-86. John Benjamins Publishing.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid. In N. Schmitt (Eds.) *Formulaic sequences: Acquisition, processing and use*, 127-151. John Benjamins Publishing.

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.

Schraw, G., Trathen, W., Reynolds, R. E., & Lapan, R. T. (1988). Preferences for idioms: Restrictions due to lexicalization and familiarity. *Journal of Psycholinguistic Research, 17*(5), 413-424.

Schweigert, W. A. (1986). The comprehension of familiar and less familiar idioms. *Journal of Psycholinguistic Research*, *15*(1), 33-45.

Schweigert, W. A. (1991). The muddy waters of idiom comprehension. *Journal of Psycholinguistic Research*, *20*(4), 305-314.

Schweigert, W. A., & Moates, D. R. (1988). Familiar idiom comprehension. *Journal of Psycholinguistic Research*, *17*(4), 281-296.

Schweigert, W. A., Cintron, J., Sullivan, K., Ilic, E., Ellis, S., Dobrowits, C., & Roberts, C. (2003). Novel figurative phrases and idioms: Phrase characteristics over multiple presentations. *Journal of Psycholinguistic Research*, *32*(4), 455-475.

Sela, T., Panzer, M. S., & Lavidor, M. (2017). Divergent and convergent hemispheric processes in idiom comprehension: The role of idioms predictability. *Journal of Neurolinguistics*, *44*, 134-146.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming?. *Journal of Experimental Psychology: Learning, memory, and cognition*, *18*(6), 1191.

Shi, Z. X., Wang, C. Q., & Zhang, J. Z. (2006). *Dictionary of Chinese idioms* [J.] 汉英成语词典. China Translation and Publishing Corporation.

Siyanova-Chanturia, A. (2015). On the 'holistic'nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, *11*(2), 285-301.

Siyanova-Chanturia, A., & Sidtis, D. V. L. (2018). What Online Processing Tells us About Formulaic Language. *Understanding Formulaic Language: A Second Language Acquisition Perspective*

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.

Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., & Van Heuven, W. J. (2017). Representation and processing of multi-word expressions in the brain. *Brain and language*, *175*, 111-122.

Sorhus, H. B. (1977). To hear ourselves—Implications for teaching English as a second language. *ELT Journal*, *31*(3), 211-221.

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*, 191-226. John Benjamins Publishing.

Sprenger, S. A., Levelt, W. J., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, *54*(2), 161-184.

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, *18*(5), 523-534.

Tabossi, P., & Zardon, F. (1993). The activation of idiomatic meaning in spoken language comprehension. In C. Cacciari and P. Tabossi, P (Eds.) *Idioms: Processing, structure, and interpretation*, 145-162. Psychology Press.

Tabossi, P., Fanari, R., & Wolf, K. (2005). Spoken idiom recognition: Meaning retrieval and word expectancy. *Journal of Psycholinguistic Research*, *34*(5), 465-495.

Tabossi, P., Fanari, R., & Wolf, K. (2008). Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 313.

Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast?. *Memory & Cognition*, *37*(4), 529-540.

Tabossi, P., Wolf, K., & Koterle, S. (2009). Idiom syntax: Idiosyncratic or principled?. *Journal of Memory and Language*, *61*(1), 77-96.

Tagliaferri, B. 2008. Paradigm: Perception Research Systems [Computer Program]. Retrieved March 23, 2008, from http://www.perceptionresearchsystems.com/.

Terrell, T. D. (1991). The role of grammar instruction in a communicative approach. *The Modern Language Journal*, *75*(1), 52-63.

Thibodeau, P., & Durgin, F. H. (2008). Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language*, *58*(2), 521-540.

Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol*, *9*(4), 247-270.

Titone, D. A., & Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of pragmatics*, *31*(12), 1655-1674.

Titone, D., & Libben, M. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon*, *9*(3), 473-496.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Eds.) *Perspectives on formulaic language: Acquisition and communication*, 151-173. Bloomsbury Publishing.

Türker, E. (2016). Idiom acquisition by second language learners: The influence of cross-linguistic similarity and context. *The Language Learning Journal*, 1-12.

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

Vaid, J. (2000). New approaches to conceptual representations in bilingual memory: The case for studying humor interpretation. *Bilingualism: Language and Cognition, 3*(1), 28-30.

Van de Voort, M. E., & Vonk, W. (1995). You Don't Die Immediately When You Kick an Empty Bucket: A Processing View on Semantic and Syntactic Characteristics of ldioms. *Idioms: Structural and psychological perspectives*, *283*.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: a practical approach to modelling cognitive*. Academic Press, London.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. Journal of Cognitive Neuroscience, 22(8), 1682-1700.

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, *16*(2), 180-205.

Weinert, R. (2010). Formulaicity and usage-based language: Linguistic, psycholinguistic and acquisitional manifestations. In D. Wood (Eds.) *Perspectives on formulaic language: Acquisition and communication,* 1-20. Bloomsbury Publishing.

Wen, D. Z. 温端政. (2007) The rise of the dictionary and its contribution to cultural heritage [J.] 语典的兴起及其对文化传承的贡献 *Studies of Dictionaries* 辞书研究, 14-28.

Wen, D. Z. 温端政. (1989) *The Chinese dictionary of colloquial language* [J.] 中国俗语大辞典, Shanghai Lexicographic Publishing House.

Whalen, K., & Menard, N. (1995). L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing. *Language learning*, *45*(3), 381-418.

Xiao, Z. S. 肖竹声 (1987) Two statistics of four-character idioms [J.] 四言成语的两项小统计, *Chinese Language World* 中国语文天地 Vol.5.

Xun, E., Rao, G., Xiao, X., Zang, J. (2015) The construction of the BCC Corpus in the age of Big Data [J.] 大数据背景下 BCC 语料库的研制 *Corpus Linguistics* 语料库语言学 (3), 93-119.

Yamashita, J. (2018). Possibility of semantic involvement in the L1-L2 congruency effect in the processing of L2 collocations. *Journal of Second Language Studies*, *1*(1), 60-78.

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.

Yang, J., Li, P., Fang, X., Shu, H., Liu, Y., & Chen, L. (2016). Hemispheric involvement in the processing of Chinese idioms: An fMRI study. *Neuropsychologia*, *87*, 12-24.

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, *13*(2), 48-67.

Zempleni, M. Z., Haverkort, M., Renken, R., & Stowe, L. A. (2007). Evidence for bilateral involvement in idiom comprehension: An fMRI study. *Neuroimage*, *34*(3), 1280-1291.

Zhang, H., Yang, Y., Gu, J., & Ji, F. (2013). ERP correlates of compositionality in Chinese idiom comprehension. *Journal of Neurolinguistics*, *26*(1), 89-112.

Zhou, J. 周荐. (1998). New perspectives on Guanyongyu [J.] 惯用语新论. *Language Teaching and Research* 语言教学与研究, (1), 128-139.

APPENDIX A: CHINESE INSTRUCTIONS FOR STUDY 1

a): Chinese instructions for familiarity 词语使用频率打分

在这个任务中，您将读到一些词语，请用 1~5 分给每个词语的使用频率打分。
1=从来没见过、听过、或者用过这个词语；5=常常见到、听到或者用到。

b) : Chinese instructions for meaningfulness 词语熟识度打分

在这个任务中， 您将读到一些词语，请用 1~5 分给您对每个词语的熟识程度打分。1=完全不认识；5=完全认识，并且能够清楚地说出它的意思。

c): Chinese instructions for compositionality 词语可分析度打分

在这个任务里，您将读到一些词语。请用 1~5 分给每个词语的可分析度打分。可分析的词语即，词语的整体含义跟每个字的字面有紧密关联。不可分析的词语即，词语的整体含义跟每个字的字面意思都没有关联。1=不可分析；5=完全可以分析。例如：“春夏秋冬”的整体含义是“一年四季”，其整体意义跟每个字的意义都紧密相关，因此是一个完全可分析的词语，得 5 分。“一石二鸟”的整体含义是“一个举动达到两个目的”，只跟部分字，比如“一”和“二”，有关联，因此是一个部分可分析的词语，得 3 分。“阳春白雪”的整体含义是“高深的文艺作品”，跟每个都没有关联，因此是一个不可分析的词语，得 1 分。

d): Chinese instructions for literality 词语字面意思现实性打分

在这个任务里，您将读到一些词语。请用 1~5 分给每个词语的字面意思在现实语境中使用的可能性打分。词语的字面意思就是每个字的字面意思合起来所表达的意思。1=字面意思完全不可能实现（不会被用到）；5=字面意思可以实现（常被用到）。例如：“叶公好龙”的字面意思是 “叶公喜欢龙”，然而现实语境中可能不存在一个人叫“叶公”，也不存在“龙”这个事物，所以“叶公好龙”的字面意思在现实语境中被使用的可能性就很低，得 1 分。“波涛汹涌”的字面意思是“波浪翻滚很剧烈”，它的字面意很可能也常常被使用在现实语境中，例如，“海面波涛汹涌”，得 5 分。

e): Chinese instructions for predictability 末尾字完型填空

在这个任务中，您将会看到一些不完整的词语，它们的末尾都缺失了一个字，需要您补全。请把您脑海中想到的第一个字填在横线上，使这个词语片段变成一个完整有意义的词语。例如：船到桥_____，您可能第一个想到的字是“头”，那么请把“头”字输入横线处。 注意:只能填写一个字。

f): Chinese instructions for linguistic register 词语语体归属判断

请根据您的语感判断，下面词语属于哪种书面（正式）语体，还是口语（非正式）语体？请勾选相应的答案。

APPENDIX B: AVERAGE RAINGS FOR 182 GYYS ON SIX DIMENSIONS

	Familiarity	Meaningfulness	Compositionality	Literality	Predictability	Register
一刀切	4.500	4.855	3.610	3.571	0.571	0.471
一把手	4.691	4.898	2.902	3.617	0.231	0.294
一面倒	4.385	4.523	3.458	3.729	0.333	0.431
万金油	3.703	4.289	3.034	3.024	0.571	0.627
三不知	4.340	4.472	3.134	3.431	0.083	0.784
不人道	4.552	4.744	3.857	3.953	0.455	0.549
不失为	3.561	3.957	2.789	2.776	0.200	0.784
不尽然	3.534	4.444	3.681	3.581	0.273	0.647
不得了	4.875	4.907	3.543	4.000	0.800	0.176
不得已	4.769	4.600	4.034	3.833	0.667	0.235
不敢当	4.630	4.826	4.314	3.934	0.538	0.235
不经意	4.846	4.631	3.525	4.167	0.833	0.490
不自量	4.052	4.711	3.958	3.744	1.000	0.667
不见得	4.638	4.756	3.647	3.814	0.636	0.098
不足道	4.141	4.446	3.254	3.313	0.167	0.569
二百五	4.828	4.778	2.504	3.791	0.636	0.275
传声筒	3.691	4.784	3.667	3.489	0.615	0.569
伤脑筋	4.603	4.822	3.899	3.977	0.818	0.275
做人情	4.217	4.609	3.843	3.459	0.154	0.529
做文章	4.397	4.756	3.017	3.605	1.000	0.471
做生意	4.930	4.957	4.298	4.245	1.000	0.157
免不了	4.559	4.875	3.765	4.234	1.000	0.176
兜圈子	4.513	4.897	3.943	3.872	0.700	0.412
全武行	2.676	3.955	3.039	2.596	0.462	0.843
八辈子	4.361	4.389	2.976	2.941	0.917	0.471
出人命	4.641	4.783	3.898	3.381	0.071	0.373
出气筒	4.772	4.804	3.719	3.878	1.000	0.314
出洋相	4.567	4.708	2.878	3.333	0.917	0.216
出风头	4.630	4.826	3.961	3.410	0.769	0.098
分水岭	4.172	4.675	3.746	4.048	0.857	0.706
到头来	4.672	4.889	3.437	3.837	1.000	0.196
刽子手	4.093	4.556	3.183	3.176	1.000	0.706
半辈子	4.825	4.825	4.443	4.615	1.000	0.196
半边天	4.172	4.722	3.908	3.837	0.818	0.490
发神经	4.672	4.856	3.790	3.930	1.000	0.392
发脾气	4.797	4.867	3.746	3.905	1.000	0.176
叠罗汉	4.038	4.185	3.119	3.833	1.000	0.706
吃不消	4.569	4.833	3.840	4.186	0.364	0.216
吃官司	4.310	4.722	3.580	3.535	0.818	0.294
吃老本	4.338	4.830	3.157	3.851	0.769	0.294
吃豆腐	4.551	4.538	2.695	3.771	0.417	0.569

吃闲饭	4.517	4.722	3.748	3.628	1.000	0.451
吹牛皮	4.563	4.771	3.305	2.929	0.857	0.176
和稀泥	4.526	4.446	2.881	3.625	1.000	0.412
咬耳朵	4.238	4.722	3.200	3.231	1.000	0.667
哑巴亏	4.625	4.938	3.857	3.564	0.900	0.647
回马枪	3.247	3.764	2.927	3.039	0.917	0.784
坐天下	3.474	4.500	3.070	2.735	0.333	0.588
坐月子	4.782	4.723	3.322	3.542	1.000	0.235
坐江山	4.088	4.196	3.608	2.878	0.467	0.647
大不了	4.696	4.739	3.784	3.541	0.538	0.039
大不敬	4.203	4.675	3.407	3.690	0.143	0.627
大杂烩	4.328	4.700	3.966	4.302	1.000	0.392
夸海口	4.538	4.569	3.288	3.042	0.917	0.529
夹生饭	4.152	4.543	3.039	3.492	0.615	0.510
套交情	4.000	4.522	3.281	3.694	0.933	0.392
好容易	4.775	4.897	3.986	3.923	0.900	0.549
好意思	4.795	4.523	3.407	4.083	0.833	0.020
对得起	4.797	4.831	3.881	3.714	0.643	0.118
差不离	3.515	4.705	3.118	3.660	0.077	0.647
干瞪眼	4.543	4.783	4.157	3.770	1.000	0.431
开倒车	3.379	3.656	3.092	3.791	0.727	0.647
开后门	4.703	4.795	3.186	3.595	0.714	0.353
开夜车	4.516	4.819	3.797	3.786	0.786	0.588
开绿灯	4.561	4.957	3.667	4.122	0.867	0.353
弄潮儿	3.868	4.557	2.608	3.255	0.692	0.686
得人心	4.515	4.639	4.378	3.471	0.917	0.569
怪不得	4.609	4.831	3.322	3.452	1.000	0.059
恨不得	4.813	4.918	3.900	3.769	0.900	0.275
惹是非	4.330	4.639	4.171	3.863	1.000	0.549
慢半拍	4.621	4.800	4.261	4.116	1.000	0.353
打下手	4.702	4.891	3.561	3.714	0.933	0.255
打主意	4.423	4.477	3.542	3.646	0.917	0.216
打交道	4.794	4.886	3.235	3.915	0.923	0.078
打出手	3.614	4.478	2.930	3.327	0.133	0.765
打前站	3.842	4.522	3.158	3.531	0.533	0.529
打哈哈	4.103	4.477	2.831	3.354	0.250	0.569
打哑谜	4.258	4.806	4.524	4.098	0.750	0.569
打嘴仗	4.422	4.904	3.847	3.143	0.143	0.451
打圆场	4.474	4.556	2.805	3.157	0.833	0.353
打埋伏	4.174	4.783	3.392	3.148	0.923	0.451
打天下	4.144	4.556	4.378	3.176	0.667	0.627
打头阵	4.462	4.569	3.542	3.896	0.750	0.275
打官司	4.825	4.652	3.982	3.959	1.000	0.353
打官腔	4.667	4.978	3.825	3.551	0.267	0.471
打寒战	4.196	4.458	2.927	3.235	1.000	0.569

打折扣	4.603	4.898	3.725	3.915	0.923	0.176
打板子	3.754	3.608	2.043	3.796	0.733	0.686
打棍子	2.887	4.609	3.754	2.627	0.750	0.725
打牙祭	3.700	4.784	2.886	2.744	0.700	0.627
打秋风	3.150	4.913	3.211	2.051	0.300	0.824
打算盘	4.269	4.569	3.407	3.875	0.917	0.412
打通关	4.299	4.458	3.427	3.686	0.417	0.667
扣帽子	4.109	4.783	3.034	3.310	1.000	0.510
执牛耳	2.957	4.196	2.490	2.311	0.615	0.922
护犊子	4.577	4.583	3.610	3.588	0.833	0.392
抱不平	4.672	4.819	3.508	3.333	0.286	0.529
抱佛脚	4.382	4.875	2.608	3.319	1.000	0.431
拉下水	4.268	4.681	3.220	3.510	0.167	0.333
拉下马	4.217	4.804	4.078	3.902	0.385	0.235
拍胸脯	4.672	4.831	3.475	3.643	0.857	0.490
拍马屁	4.821	4.708	3.068	3.521	1.000	0.118
拜天地	4.138	4.722	3.580	3.907	1.000	0.392
拿主意	4.825	4.739	4.228	3.939	1.000	0.294
挑大梁	4.484	4.819	3.729	3.452	0.786	0.529
挖墙脚	4.529	4.864	2.706	3.660	1.000	0.157
挡箭牌	4.587	4.913	4.471	4.262	1.000	0.490
挨板子	4.370	4.739	4.275	4.164	1.000	0.549
换脑筋	4.088	4.773	3.486	2.795	0.200	0.569
掏腰包	4.426	4.864	3.333	4.234	1.000	0.373
撒手锏	3.691	4.545	2.647	3.277	0.385	0.569
撒酒疯	4.672	4.880	4.119	3.690	0.786	0.314
敲竹杠	4.034	4.611	2.571	3.558	0.364	0.392
断头台	4.275	4.897	4.571	3.538	0.700	0.627
无底洞	4.448	4.889	4.328	3.953	0.727	0.373
暗地里	4.588	4.928	4.086	3.564	1.000	0.275
来不及	4.900	4.948	4.429	4.487	0.800	0.275
架不住	4.087	4.739	4.098	3.574	0.692	0.490
泡病号	3.109	4.109	3.020	2.607	0.385	0.588
泥饭碗	3.123	4.348	2.895	2.367	0.733	0.588
滚雪球	4.565	4.870	3.804	4.016	1.000	0.549
满天飞	4.587	4.587	3.804	4.033	0.154	0.725
煞风景	4.423	4.631	3.237	3.458	0.917	0.294
爬格子	3.062	3.611	2.402	2.549	0.833	0.745
犯不上	4.672	4.722	3.429	3.651	0.636	0.431
犯嘀咕	4.474	4.891	3.298	3.653	1.000	0.373
狗腿子	4.350	4.887	3.014	3.103	1.000	0.431
留尾巴	3.789	4.652	3.211	3.306	0.733	0.510
百事通	4.600	4.876	4.314	4.000	0.200	0.745
皮包骨	4.474	4.597	4.012	3.882	0.833	0.549
看不起	4.781	4.843	3.898	3.833	0.500	0.137

看得起	4.772	4.933	4.018	3.959	0.067	0.235
看热闹	4.765	4.864	4.078	4.681	0.923	0.137
破天荒	4.638	4.856	3.514	2.897	0.900	0.451
砸饭碗	4.261	4.913	3.961	3.836	0.923	0.275
硬碰硬	4.526	4.764	3.902	3.765	1.000	0.333
碰钉子	4.625	4.918	3.786	3.590	1.000	0.333
禁不住	4.575	4.835	3.971	4.359	0.800	0.431
禁不起	4.259	4.522	3.824	3.837	0.455	0.510
空对空	3.326	4.370	3.137	2.689	0.154	0.804
窝里斗	4.609	4.826	4.235	3.836	0.692	0.373
等不及	4.813	4.928	4.571	4.410	0.700	0.235
绊脚石	4.397	4.886	3.686	4.191	0.615	0.196
绕圈子	4.700	4.948	4.157	4.103	0.700	0.451
绞脑汁	4.192	4.508	3.220	2.625	0.750	0.529
翻白眼	4.754	4.935	4.140	4.510	1.000	0.176
耳边风	4.731	4.723	3.373	3.729	0.750	0.235
背包袱	4.059	4.841	3.000	3.872	0.231	0.549
莫不是	3.814	4.292	3.256	3.353	0.333	0.725
莫须有	4.234	4.783	3.424	3.429	0.929	0.824
装门面	4.613	4.845	3.886	3.744	0.500	0.549
要面子	4.676	4.943	3.294	4.128	0.923	0.078
见光死	4.063	4.687	3.542	2.857	0.571	0.588
豁出去	4.692	4.723	3.712	3.833	1.000	0.157
走后门	4.797	4.855	3.288	3.929	0.929	0.275
走过场	4.672	4.822	3.555	4.000	0.364	0.431
走钢丝	4.034	4.689	2.731	3.372	1.000	0.706
赶时髦	4.842	4.630	4.018	4.082	0.733	0.353
赶浪头	2.522	4.109	2.667	2.934	0.000	0.804
跑龙套	4.603	4.600	2.797	3.479	1.000	0.471
软脚蟹	3.128	3.354	3.034	2.938	0.250	0.608
过不去	4.474	4.597	3.793	3.804	0.667	0.157
过得去	4.536	4.653	3.634	3.824	0.333	0.216
过日子	4.897	4.723	3.746	4.354	0.917	0.196
进一步	4.863	4.887	4.557	4.385	0.700	0.451
避风头	4.132	4.864	3.157	4.064	0.231	0.490
钻空子	4.761	4.826	4.078	4.098	1.000	0.314
铁三角	4.406	4.855	3.339	3.690	0.857	0.627
铁公鸡	4.630	4.870	3.412	3.328	1.000	0.294
铁饭碗	4.544	4.875	2.745	3.681	0.769	0.353
闹洞房	4.478	4.913	4.471	4.230	1.000	0.490
降半旗	4.261	4.696	4.157	3.656	0.385	0.549
难为情	4.603	4.886	3.569	3.681	0.923	0.216
集大成	3.375	4.629	3.471	3.436	0.900	0.843
露头角	4.788	4.856	3.771	3.154	0.700	0.647
靠不住	4.691	4.830	3.627	4.511	0.923	0.137

靠得住	4.930	4.913	4.439	4.388	0.933	0.196
面面观	2.529	4.273	2.902	3.021	0.154	0.745
飞毛腿	4.044	4.841	2.745	2.979	0.846	0.275
骨子里	4.750	4.856	3.957	3.590	0.700	0.216
鬼门关	4.500	4.848	4.157	3.426	0.769	0.510
黑吃黑	4.283	4.717	3.529	3.164	0.692	0.471

APPENDIX C: AVERAGE RAINGS FOR 243 CYS ON SIX DIMENSIONS

	Familiarity	Meaningfulness	Compositionality	Literality	Predictability	Register
一丝不挂	4.276	4.911	3.807	3.884	0.545	0.667
一丝不苟	4.641	4.708	3.949	3.771	0.833	0.963
一厢情愿	4.848	4.891	4.314	3.787	1.000	0.778
一如既往	4.703	4.855	3.814	3.667	1.000	0.741
一帆风顺	4.765	4.932	4.216	4.255	1.000	0.741
一席之地	4.596	4.804	4.211	3.735	1.000	0.741
一心一意	4.711	4.833	4.646	3.980	1.000	0.407
一成不变	4.655	4.833	4.311	3.930	0.727	0.778
一无所有	4.789	4.913	4.386	4.367	0.667	0.519
一模一样	4.838	4.959	4.714	4.667	1.000	0.593
一步到位	4.559	4.920	4.176	4.298	1.000	0.519
一目了然	4.783	4.935	4.765	4.410	1.000	0.741
一见钟情	4.672	4.880	4.339	4.000	1.000	0.556
一视同仁	4.724	4.822	3.790	3.581	1.000	0.815
一触即发	4.441	4.932	4.118	4.170	0.923	0.963
一针见血	4.629	4.750	3.524	3.745	1.000	0.852
不亦乐乎	4.725	4.866	3.800	3.744	1.000	0.778
不以为然	4.632	4.818	3.510	4.106	0.769	0.556
不可思议	4.732	4.764	4.159	3.745	0.833	0.667
不可或缺	4.667	4.492	3.831	3.917	1.000	0.852
不可收拾	4.522	4.826	4.333	4.016	0.538	0.556
不折不扣	4.691	4.653	3.524	3.608	0.833	0.667
不择手段	4.670	4.806	4.012	3.745	1.000	0.815
不正之风	4.603	4.615	4.254	3.688	0.500	0.741
不由自主	4.629	4.875	4.000	4.039	0.917	0.704
不知不觉	4.735	4.875	4.333	4.660	0.692	0.407
不知所云	4.431	4.822	4.076	3.953	0.500	0.778
不知所措	4.647	4.830	3.980	4.383	0.846	0.630
不约而同	4.756	4.708	4.525	4.229	1.000	0.667
不言而喻	4.577	4.615	4.441	4.125	1.000	0.815
与生俱来	4.788	4.907	4.529	4.333	1.000	0.815
丰富多彩	4.759	4.900	4.521	3.837	0.909	0.815
举足轻重	4.667	4.554	3.085	3.417	1.000	0.667
义无反顾	4.702	4.696	3.789	3.327	1.000	0.852
乐此不疲	4.546	4.472	4.110	3.706	1.000	0.815
五花八门	4.707	4.856	3.504	3.465	1.000	0.741
五颜六色	4.860	4.913	4.649	4.551	0.867	0.667
人山人海	4.618	4.943	4.059	4.106	1.000	0.667
以身作则	4.863	4.887	4.386	4.462	0.900	0.778
众所周知	4.813	4.938	4.571	4.308	1.000	0.778
何去何从	4.667	4.692	4.102	3.979	1.000	0.519

供不应求	4.804	4.891	4.725	4.590	1.000	0.630
全力以赴	4.794	4.597	3.073	3.235	1.000	0.778
全心全意	4.574	4.864	4.373	4.574	1.000	0.704
兴高采烈	4.707	4.844	4.218	3.977	1.000	0.778
再接再厉	4.930	4.848	4.246	3.898	1.000	0.815
冰天雪地	4.630	4.891	4.784	4.426	1.000	0.704
出人意料	4.754	4.804	4.386	4.327	0.933	0.519
出神入化	4.667	4.609	3.491	3.245	0.933	0.926
出类拔萃	4.717	4.870	4.235	3.672	1.000	0.889
出谋划策	4.574	4.875	4.333	4.340	1.000	0.704
初来乍到	4.338	4.875	4.137	4.340	1.000	0.556
别具一格	4.563	4.783	3.508	3.476	1.000	0.889
别出心裁	4.586	4.833	3.387	3.349	0.909	0.815
刮目相看	4.756	4.708	3.254	3.521	1.000	0.778
刻不容缓	4.804	4.891	4.294	4.295	1.000	0.963
刻骨铭心	4.679	4.738	4.322	3.375	1.000	0.963
前所未有	4.529	4.818	4.431	4.319	1.000	0.852
力不从心	4.800	4.948	4.486	4.436	1.000	0.704
力所能及	4.724	4.844	4.361	4.070	1.000	0.667
勇往直前	4.722	4.819	4.598	4.392	1.000	0.704
匪夷所思	4.578	4.795	3.356	3.405	1.000	0.926
千奇百怪	4.731	4.692	4.407	3.729	1.000	0.778
千方百计	4.670	4.819	4.293	3.902	1.000	0.815
千篇一律	4.789	4.913	4.421	4.082	0.800	0.815
博大精深	4.744	4.677	3.932	4.250	1.000	0.852
卷土重来	4.700	4.938	3.986	3.256	1.000	0.778
原汁原味	4.825	4.739	4.439	4.551	1.000	0.593
反腐倡廉	4.737	4.870	4.737	4.571	0.867	0.630
发扬光大	4.618	4.909	4.020	4.106	1.000	0.889
取而代之	4.789	4.913	4.579	4.531	1.000	0.852
古色古香	4.422	4.855	4.237	3.905	1.000	0.741
可想而知	4.647	4.886	4.137	4.574	1.000	0.481
叹为观止	4.074	4.716	3.431	3.766	0.923	0.815
各式各样	4.875	4.948	4.657	4.718	1.000	0.741
同舟共济	4.800	4.938	4.286	3.897	1.000	0.963
名列前茅	4.706	4.830	3.549	3.872	1.000	0.815
名副其实	4.850	4.918	4.200	4.103	1.000	0.889
后顾之忧	4.750	4.897	4.200	3.974	1.000	0.815
呼之欲出	4.544	4.891	4.105	3.673	0.667	0.704
咬牙切齿	4.505	4.708	3.768	4.020	1.000	0.704
哭笑不得	4.782	4.662	4.119	4.125	1.000	0.444
因人而异	4.782	4.677	4.322	4.417	1.000	0.778
因地制宜	4.641	4.615	3.712	4.042	1.000	0.667
图文并茂	4.649	4.826	4.386	4.429	0.800	0.778
多愁善感	4.860	4.978	4.456	4.367	1.000	0.852

大千世界	4.630	4.804	4.255	3.246	0.769	0.704
大吃一惊	4.850	4.938	4.257	3.615	1.000	0.444
大同小异	4.848	4.913	4.765	3.902	1.000	0.741
大惊小怪	4.676	4.875	4.000	3.894	1.000	0.704
天人合一	4.203	4.675	3.542	3.500	0.857	0.815
天南地北	4.586	4.822	4.134	3.372	1.000	0.741
天涯海角	4.598	4.764	4.671	4.308	1.000	0.667
天马行空	4.738	4.918	3.543	2.897	1.000	0.852
如火如荼	4.391	4.761	3.922	3.459	1.000	0.852
学以致用	4.588	4.722	4.354	4.196	0.917	0.778
安居乐业	4.706	4.932	4.412	4.319	1.000	0.704
家喻户晓	4.763	4.778	4.659	3.980	1.000	0.815
对症下药	4.662	4.909	4.294	4.532	1.000	0.815
小心翼翼	4.759	4.844	3.924	4.000	1.000	0.704
尘埃落定	4.464	4.639	3.512	3.451	0.833	0.963
尽如人意	4.526	4.538	4.169	4.021	1.000	0.630
层出不穷	4.531	4.867	3.203	3.405	0.857	0.815
应有尽有	4.734	4.867	4.525	4.333	1.000	0.667
异口同声	4.842	4.957	4.596	4.551	0.867	0.852
弄虚作假	4.742	4.681	4.720	4.137	1.000	0.741
引人入胜	4.449	4.431	3.559	3.729	1.000	0.889
当务之急	4.544	4.886	3.902	4.128	1.000	0.852
形形色色	4.703	4.819	3.729	3.976	1.000	0.704
得不偿失	4.649	4.736	4.134	3.902	0.917	0.852
得天独厚	4.456	4.841	3.510	3.766	0.923	0.889
得心应手	4.707	4.822	3.908	3.814	1.000	0.778
循序渐进	4.567	4.667	4.305	3.804	0.917	0.852
微不足道	4.863	4.948	4.343	4.231	1.000	0.852
心不在焉	4.734	4.819	4.102	3.786	1.000	0.556
心中有数	4.776	4.844	4.235	4.000	0.545	0.407
心平气和	4.632	4.909	4.294	4.553	1.000	0.704
心心相印	4.275	4.897	4.357	3.333	0.800	0.778
心旷神怡	4.615	4.538	4.085	4.063	1.000	0.963
心满意足	4.828	4.771	4.492	4.167	1.000	0.519
心甘情愿	4.753	4.833	4.671	3.902	1.000	0.630
必由之路	4.062	4.319	4.195	3.745	0.917	0.704
志同道合	4.813	4.938	4.600	4.205	1.000	0.889
念念不忘	4.647	4.886	3.843	4.574	1.000	0.630
急功近利	4.638	4.767	4.235	3.744	0.818	0.852
总而言之	4.639	4.681	3.988	4.020	1.000	0.741
恰到好处	4.825	4.870	4.561	4.327	1.000	0.593
恶性循环	4.588	4.920	4.314	4.426	1.000	0.667
情不自禁	4.703	4.843	3.983	4.119	1.000	0.778
情有独钟	4.679	4.646	3.847	3.750	1.000	0.741
惊天动地	4.660	4.750	4.268	3.843	1.000	0.815

惊心动魄	4.456	4.886	4.196	4.170	0.923	0.630
惨不忍睹	4.625	4.855	4.153	4.238	0.929	0.667
愈演愈烈	4.483	4.644	4.176	3.628	0.818	0.593
成千上万	4.750	4.909	4.373	4.532	0.923	0.704
我行我素	4.731	4.662	3.915	3.854	1.000	0.815
扑朔迷离	4.443	4.639	2.805	2.804	1.000	0.852
持之以恒	4.603	4.733	4.101	3.860	1.000	0.852
排忧解难	4.536	4.681	4.646	4.137	0.833	0.815
推陈出新	4.391	4.783	4.373	4.190	0.929	0.963
无动于衷	4.731	4.646	3.746	3.833	1.000	0.815
无可厚非	4.469	4.711	2.881	3.333	0.786	0.926
无可奈何	4.638	4.811	3.815	3.907	1.000	0.556
无怨无悔	4.782	4.785	4.627	4.188	0.917	0.889
无所事事	4.756	4.708	3.983	4.292	1.000	0.593
无所适从	4.543	4.848	4.020	3.656	0.769	0.741
无济于事	4.696	4.848	4.412	3.689	1.000	0.815
无能为力	4.848	4.935	4.569	4.279	1.000	0.593
无论如何	4.759	4.856	4.050	3.674	0.909	0.667
日新月异	4.688	4.876	4.200	3.436	1.000	0.852
昙花一现	4.475	4.959	3.843	3.615	1.000	0.926
有声有色	4.588	4.611	3.793	3.980	0.917	0.630
有朝一日	4.531	4.807	3.847	3.595	1.000	0.815
有条不紊	4.670	4.722	3.988	3.627	1.000	0.815
有的放矢	4.543	4.522	3.235	3.328	0.538	0.889
有目共睹	4.826	4.870	4.667	4.279	0.769	0.741
有识之士	4.345	4.822	4.067	3.837	1.000	0.778
未雨绸缪	4.500	4.663	3.593	3.595	0.929	0.889
标本兼治	4.256	4.898	3.898	3.667	0.750	0.963
根深蒂固	4.719	4.957	4.474	4.102	0.867	0.889
格格不入	4.696	4.848	4.020	3.475	1.000	0.741
梦寐以求	4.680	4.722	4.171	3.922	1.000	0.889
欢天喜地	4.930	4.978	4.439	3.837	1.000	0.704
歇斯底里	4.649	4.630	2.772	3.265	0.867	0.741
此起彼伏	4.569	4.800	4.336	3.837	0.909	0.704
求真务实	4.630	4.826	4.686	4.016	0.769	0.519
沸沸扬扬	4.789	4.870	3.860	3.510	1.000	0.667
淋漓尽致	4.826	4.848	4.196	3.377	1.000	0.778
游刃有余	4.663	4.866	3.686	3.385	1.000	0.741
游山玩水	4.912	4.957	4.737	4.571	1.000	0.519
源远流长	4.250	4.773	3.980	3.979	1.000	0.815
漫不经心	4.667	4.646	3.831	3.771	1.000	0.741
潜移默化	4.456	4.830	3.392	4.149	1.000	0.926
炙手可热	4.654	4.585	3.254	3.104	0.917	0.963
热血沸腾	4.594	4.843	4.119	3.452	1.000	0.481
焕然一新	4.670	4.764	4.134	3.863	1.000	0.963

爱不释手	4.629	4.750	4.183	3.961	1.000	0.630
独一无二	4.649	4.833	4.634	4.078	1.000	0.778
独树一帜	4.652	4.783	4.294	3.705	0.769	0.852
独立自主	4.756	4.662	4.407	4.229	1.000	0.630
理所当然	4.877	4.891	4.351	4.367	1.000	0.815
理直气壮	4.750	4.771	4.271	4.214	1.000	0.704
琳琅满目	4.324	4.773	3.765	4.000	1.000	0.852
甜言蜜语	4.828	4.889	4.521	4.186	1.000	0.556
畅所欲言	4.680	4.681	4.415	3.686	0.917	0.852
白手起家	4.656	4.807	3.831	3.595	1.000	0.630
百家争鸣	4.663	4.938	4.143	3.564	0.800	0.704
百花齐放	4.515	4.778	3.988	3.647	1.000	0.778
目瞪口呆	4.691	4.806	4.305	3.922	1.000	0.815
眼花缭乱	4.652	4.891	4.627	3.984	1.000	0.704
知己知彼	4.826	4.935	4.804	4.492	0.923	0.815
突如其来	4.703	4.819	3.983	3.714	1.000	0.889
突飞猛进	4.701	4.722	4.280	3.863	1.000	0.778
精打细算	4.782	4.692	4.322	4.125	1.000	0.778
精益求精	4.652	4.848	4.608	3.918	0.769	0.741
紧锣密鼓	4.219	4.771	3.424	3.381	0.857	0.963
耐人寻味	4.382	4.773	3.608	3.809	1.000	0.815
耳熟能详	4.485	4.773	4.196	4.298	0.846	0.926
耳目一新	4.547	4.795	3.814	3.357	0.857	0.667
胡说八道	4.859	4.708	3.847	4.104	1.000	0.296
脚踏实地	4.808	4.692	4.424	4.188	0.833	0.741
脱口而出	4.741	4.833	4.168	4.209	1.000	0.704
脱颖而出	4.782	4.646	3.797	3.833	1.000	0.889
自以为是	4.618	4.920	4.039	4.340	0.923	0.667
自始至终	4.632	4.920	4.431	4.809	1.000	0.778
自强不息	4.641	4.819	4.288	4.167	1.000	0.815
自欺欺人	4.691	4.898	4.216	4.638	1.000	0.852
自然而然	4.656	4.843	3.729	3.786	1.000	0.481
自言自语	4.850	4.959	4.686	4.641	1.000	0.667
舞文弄墨	4.152	4.761	3.784	3.426	0.769	0.852
花花世界	4.485	4.886	3.176	3.574	1.000	0.704
若无其事	4.807	4.935	4.386	4.408	1.000	0.778
若隐若现	4.667	4.783	4.561	4.224	1.000	0.889
茶余饭后	4.441	4.841	3.922	4.319	1.000	0.593
莫名其妙	4.662	4.886	3.784	4.064	1.000	0.593
蒸蒸日上	4.789	4.870	3.719	3.306	1.000	0.889
蠢蠢欲动	4.684	4.978	4.000	3.694	0.867	0.741
见义勇为	4.910	4.692	4.305	4.396	1.000	0.926
触目惊心	4.567	4.708	4.390	3.824	1.000	0.889
讨价还价	4.814	4.833	4.293	4.471	1.000	0.407
谈天说地	4.632	4.935	4.439	3.918	0.867	0.704

赏心悦目	4.739	4.826	4.784	4.066	1.000	0.852
足不出户	4.763	4.948	4.557	4.333	0.800	0.667
软硬兼施	4.609	4.783	4.686	3.885	0.615	0.630
轰轰烈烈	4.795	4.692	4.034	3.729	1.000	0.704
轻而易举	4.848	4.935	4.373	4.279	1.000	0.741
迎刃而解	4.441	4.818	3.549	3.830	0.923	0.889
运筹帷幄	4.491	4.826	3.298	3.245	1.000	0.852
迫不及待	4.647	4.932	4.059	4.426	1.000	0.778
迫在眉睫	4.578	4.819	3.661	3.071	0.786	0.963
铺天盖地	4.621	4.856	4.109	3.628	1.000	0.704
错综复杂	4.448	4.689	4.412	4.093	1.000	0.815
锦上添花	4.594	4.747	3.780	4.214	1.000	0.963
雪上加霜	4.763	4.938	4.129	3.872	1.000	0.852
雷霆万钧	4.200	4.670	3.643	3.179	1.000	0.852
青梅竹马	4.795	4.723	2.966	3.250	1.000	0.704
面目全非	4.588	4.909	4.078	4.191	1.000	0.815
顾名思义	4.800	4.918	4.286	3.949	1.000	0.889
风花雪月	4.674	4.761	3.059	3.344	0.769	0.741
风起云涌	4.475	4.856	3.957	3.923	1.000	0.926
高高在上	4.772	4.957	4.526	4.265	0.867	0.444
默默无闻	4.588	4.898	4.118	4.255	1.000	0.815
齐心协力	4.735	4.852	4.647	4.340	1.000	0.852

APPENDIX D: TEST MATERIALS FOR STUDY 2

GYY

Related idiom	Related novel phrase	Unrelated idiom	Target
*走钢丝	切钢板	不自量	鐵匠
敲竹杠	撑竹竿	做人情	熊貓
飞毛腿	短毛狗	吃老本	皮囊
背包袱	背包客	避风头	行李
见光死	遮光板	大杂烩	照明
扣帽子	戴帽子	打埋伏	遮陽
分水岭	洒水车	大不敬	飲用
半边天	左边走	耳边风	疆界
狗腿子	猪腿肉	铁公鸡	走路
抱佛脚	摆佛像	打哑谜	教徒
绊脚石	洗脚盆	断头台	步伐
掏腰包	系腰带	慢半拍	脊椎
无底洞	脚底板	铁三角	限制
鬼门关	大门外	骨子里	通過
*挖墙脚	靠墙站	看热闹	围欄
夸海口	去海边	打圆场	水面
铁饭碗	电饭锅	拍胸脯	食物
开绿灯	穿绿衣	暗地里	植物
吹牛皮	赶牛车	吃官司	馬匹
挡箭牌	神箭手	闹洞房	靶子
破天荒	全天下	怪不得	空間
跑龙套	穿龙袍	吃豆腐	傳說
伤脑筋	费脑力	难为情	神經
碰钉子	拔钉子	兜圈子	螺絲
出风头	看风景	发神经	浪潮
*百事通	有事吗	窝里斗	辦理
打官腔	当官的	要面子	職務
撒酒疯	摔酒瓶	滚雪球	宴席
翻白眼	穿白鞋	露头角	色彩
出气筒	出气口	半辈子	喘息

CY

Related idiom	Related novel phrase	Unrelated idiom	Target
青梅竹马	喝梅子酒	软硬兼施	花瓣
风花雪月	卖花姑娘	勇往直前	園林
全力以赴	用力敲门	不约而同	能量
*举足轻重	女足比赛	自始至终	腿腳
炙手可热	伸手去抓	与生俱来	寫字
*一针见血	被针扎到	心甘情愿	縫補
天人合一	没人看见	雪上加霜	動物
天马行空	策马而去	大同小异	拉車
未雨绸缪	多雨天气	推陈出新	灌溉

游刃有余	利刃伤人	不亦乐乎	鋒芒
因地制宜	异地恋爱	恰到好处	耕種
舞文弄墨	英文课本	必由之路	字符
有声有色	有声读物	一成不变	音樂
一丝不挂	一丝头发	恶性循环	棉花
标本兼治	人本思想	一见钟情	源頭
当务之急	公务在身	原汁原味	處理
如火如荼	生火做饭	独立自主	取暖
卷土重来	国土开发	热血沸腾	泥沙
百家争鸣	儒家经典	力所能及	學派
脱口而出	闭口不谈	耳目一新	進食
一步到位	大步向前	一帆风顺	階段
古色古香	同色领带	茶余饭后	黑白
同舟共济	龙舟比赛	各式各样	木筏
循序渐进	有序进行	无济于事	號碼
见义勇为	大义之举	咬牙切齿	仁愛
*一厢情愿	这厢有礼	家喻户晓	偏房
*刻骨铭心	接骨手术	排忧解难	架構
*以身作则	亲身示范	兴高采烈	體會
千篇一律	一篇日记	名副其实	章節
齐心协力	用心学习	精打细算	情緒

***** These items are excluded in the post hoc analysis because of a higher-than-18% error rate.

APPENDIX E: TEST MATERIALS FOR STUDY 3

3-Idiom	Log (frequency)	Stroke	3-FS	Log (frequency)	Stroke
看上去	4.214	17	看不见	4.398	17
来得及	3.948	21	来源于	4.053	23
看不起	3.610	23	看着他	3.877	25
好意思	3.234	28	好想吃	3.886	25
打交道	3.781	23	打麻将	3.799	25
过日子	3.682	13	过几天	3.810	12
不得已	3.752	18	不能做	3.815	25
不见得	3.808	19	看见你	3.909	20
看热闹	3.427	27	看一遍	3.550	22
等不及	3.496	19	等着他	3.505	28
好容易	3.290	24	好像要	3.477	28
半辈子	3.296	20	半年后	3.440	17
打官司	3.336	18	打死了	3.420	13
暗地里	3.306	26	被子里	3.375	20
不敢当	3.223	21	不敢动	3.344	21
靠得住	3.012	33	靠窗的	2.942	35
出人命	2.793	15	出不去	3.027	14
过得去	3.147	22	过多久	3.196	15
难为情	3.420	25	难接受	3.597	29
打主意	2.810	23	打桌球	2.809	16
慢半拍	2.364	27	慢下来	2.852	24
开夜车	2.281	16	别的车	2.449	19
吹牛皮	2.316	16	吹走了	2.591	16
走后门	3.017	16	走出门	3.212	15
Mean	*3.445*	*21.25*	*Mean*	*3.430*	*20.708*
4-Idiom	Log (frequency)	Stroke	4-FS	Log (frequency)	Stroke
无能为力	3.825	20	不能完全	3.742	27
一无所有	3.616	19	一定会有	3.369	21
不约而同	3.800	22	不同的是	3.742	27
迫不及待	3.907	24	看不下去	2.984	21
一见钟情	3.515	25	一段感情	2.281	34
心甘情愿	3.773	34	心爱的人	3.113	24
不可或缺	3.682	27	不可能有	3.709	25
前所未有	4.111	28	前段时间	3.108	26
不择手段	3.563	25	重要手段	3.711	31
无可奈何	4.024	24	无论如何	4.183	23
一帆风顺	3.450	20	一切顺利	2.898	21

天涯海角	3.396	32	天气真好	2.076	24
哭笑不得	3.514	35	哭了起来	3.435	29
脱口而出	3.628	25	开口说话	3.431	29
轻而易举	3.687	32	轻松愉快	3.153	29
一目了然	3.592	20	了解一下	3.113	19
出人意料	3.598	30	出门在外	2.674	19
兴高采烈	3.681	34	我很高兴	3.471	32
千方百计	4.162	17	各种方式	3.394	25
得不偿失	3.245	31	得到一个	3.284	23
大吃一惊	3.877	21	大吃一顿	2.322	20
谈天说地	2.807	29	谈论一下	2.276	20
供不应求	3.744	26	供应产品	2.969	30
心中有数	3.123	30	心里有事	2.152	25
Mean	*3.544*	*26.25*	*Mean*	*3.108*	*25.166*

APPENDIX F: COMPOSITIONALITY RATINGS FOR IDIOMS IN STUDY 3

3-Idiom	Ave. Rating	4-Idiom	Ave. Rating
看上去	3.797	无能为力	4.569
来得及	4.281	一无所有	4.386
看不起	3.898	不约而同	4.525
好意思	3.407	迫不及待	4.059
打交道	3.235	一见钟情	4.339
过日子	3.746	心甘情愿	4.671
不得已	4.034	不可或缺	3.831
不见得	3.647	前所未有	4.431
看热闹	4.078	不择手段	4.012
等不及	4.571	无可奈何	3.815
好容易	3.986	一帆风顺	4.216
半辈子	4.443	天涯海角	4.671
打官司	3.982	哭笑不得	4.119
暗地里	4.086	脱口而出	4.168
不敢当	4.314	轻而易举	4.373
靠得住	4.439	一目了然	4.765
出人命	3.898	出人意料	4.386
过得去	3.634	兴高采烈	4.218
难为情	3.569	千方百计	4.293
打主意	3.542	得不偿失	4.134
慢半拍	4.261	大吃一惊	4.257
开夜车	3.797	谈天说地	4.439
吹牛皮	3.305	供不应求	4.725
走后门	3.288	心中有数	4.235
Mean	*3.885*	*Mean*	*4.318*