Investigating the Development of Formulaic Sequences in Second Language Mandarin in a Study Abroad Context

Xianghong Sun

Submitted in accordance with the requirements for the degree of
Master of Arts by Research

The University of Leeds

School of Languages Cultures & Societies

December 2023

Acknowledgements

The completion of this thesis would have been impossible without the guidance, support and encouragement of several people to whom I would like to extend my deepest gratitude for their contribution:

First, my deep appreciation and gratitude go to my supervisors, Dr. Clare Wright and Dr. Ying Peng. During this time, you have always provided me with invaluable support, guidance and advice. Your experience and knowledge have added immense value throughout my process of completing this thesis.

Secondly, I want to express my gratitude to my family, especially my parents and sister, as well as my friends for their assistance and support during challenging times, particularly when I experienced family losses.

Finally, I owe thanks also to my companion Cosimo who has been taking care of me over the past three months, enabling me to fully focus on my writing.

Abstract

From a usage-based perspective, language comprises constructions which are 'conventionalized pairings of form and function' (Goldberg, 2006, pp. 3-5). These conventionalised expressions, often referred to as formulaic sequences (FS) (Wray, 2002). The significance of FS is evident in written texts,especially in vocabulary development.

The study abroad (SA) context can facilitate FS learning, due to the opportunity to immerse themselves in authentic target language input (Siyanova-Chanturia, 2015). However, findings remain inconsistent, and to date, very little is known about the nature and likely development of FS in L2 Chinese writing, prompting this particular study.

This research is a small-scale but in-depth mixed-method investigation into the use and development of L2 Chinese FS learning in SA setting. The research questions focus on frequency and accuracy of FS development over time for a cohort of 26 learners of Chinese who were enrolled at a UK university and the influence of FS in writing performance.

The study uses quantitative and qualitative methods to analyse frequency and accuracy in using three target types of FS in these diaries: fixed, semi-fixed, and free forms by using diaries as data resource.

The results show mixed findings, with no consistent group-wide improvement in FS quantity and accuracy over time for the participants, except for a fixed type of FS quantity and a noticeable variation among individuals. The association between the use of FS and writing output is also investigated to explore the role of FS in improving writing performance in SA context.

The findings not only illuminate the individual differences in learning FS but also shed light on the varying levels of complexity associated with different categories of FS in Mandarin, as well as the influence of FS on writing performance, which offer valuable implications for the development of teaching and learning models for Mandarin SA programmes.

Table of Contents

List of tables

List of figures

List of Charts

Chapter 1. Introduction

Background

Formulaic sequences play a crucial role in language learning, serving as chunks (Ellis, 1996; Sinclair, 1996) and a database (Ellis, 1996) that contributes significantly to vocabulary and expression learning. Recently, there has been considerable empirical interest in formulaic sequences. However, due to its broad scope covering more than forty areas such as collocations, idioms, lexical phrases, and multi-word expressions (Wray, 2002), defining and classifying formulaic sequences pose a challenge.

Building upon a thorough review of previous research, this study first defines and classifies formulaic sequences in Chinese into three types based on their fixedness: fixed, semi-fixed, and free forms. Furthermore, the study explores the contributions of formulaic sequences to L2 learning, with a specific emphasis on their impact on L2 writing skills, measured through mean sentence length and lexical diversity.

Structure of the thesis:

Chapter 2 lays the theoretical groundwork for the study, commencing with an exploration of how formulaic sequences are defined and classified. Based on the previous studies, the definition and categorization of formulaic sequences in Chinese is proposed. In this chapter, we selects three categories based on their degree of fixedness: fixed, semi-fixed, and free forms. Furthermore, it delves into the learning of formulaic sequences within a study abroad context, probing its impact on writing skills.

In Chapter 3, the focus shifts towards methodology and research design. The method of the extraction of three types of formulaic sequences are introduced which involves five main steps: preparation, extracting formulaic sequences, assessing conventionality, evaluating the correctness of formulaic sequences extracted, and checking their correlation with the writing output which are measured by lexical diversity and mean sentence length.

Chapter 4 investigate the frequency and accuracy of each type of formulaic sequence and explore the potential role these sequences can play in writing outputs, aiming a to address the research questions presented in Chapter 3.

Chapter 5 aims to answer the three research questions which are: Q1. What patterns can be seen in the usage of formulaic sequence according to their fixedness?; Q2: Did the usage of formulaic sequence, measured by frequency and accuracy, change over time, from Time 1 to Time 2?;Did the presence of the three types in of formulaic sequence (fixed; semi-fixed; free formulaic sequence) make a difference when comparing the lexical and syntax complexity of total outputs, measured by MLU and lexical diversity respectively?

Chapter 6 comes to the discussion, in which the differences of characteristics and usage patterns of all three types of formulaic sequences are discussed, aiming to explore the reasons of lower utilization of certain type of formulaic sequence and shed lights on the participants' accuracy in employing formulaic sequences.

Chapter 7 concludes the pivotal role that formulaic sequences play in language learning, especially its importance in written text. The implications and contributions of this study, as well as its limitations are discussed in this chapter. It also points out the possible suggestions for the future research.

Chapter 2 Literature Review

2.1 Introduction

In this chapter, I introduce issues concerning the definition and classification of formulaic sequences, along with their contributions to L2 learning and influences on writing output, both in general and with a specific focus on formulaic sequences in Chinese.

The first section of this chapter, Section 2.2, begins by investigating the general definition and classification of formulaic sequences. Subsequently, I delve into how formulaic sequences in Chinese are defined. Following that, I explain the reasons why, in this study, three different forms of formulaic sequences are chosen for examination, aiming to understand their usage and influence.

Section 2.3 discusses the crucial role that formulaic sequences play in L2 learning, particularly in L2 Chinese learning, and explores how a study abroad context can facilitate formulaic sequence learning. Moreover, it investigates the influence that the usage of formulaic sequences can have on writing skills, as well as the challenges associated with learning formulaic sequences.

Section 2.4 shifts its focus to explaining why diaries are chosen as the data source. Then, in Section 2.5, I provide a summary of all the relevant literature.

2.2 Formulaic sequences

'Formulaic sequence' is an umbrella term which includes over forty areas, such as collocations, idioms, lexical phrases, lexical bundles, metaphors, proverbs, phrasal verbs, n-grams, compounds and multi word expression and so on (Wray, 2002).

2.1.1 Identifying formulaic sequence

Due to the multi-faceted nature of formulaic language, it has been characterised according to its form, function, semantic, syntactic and lexical properties. In order to unify the massive terms used to indicate formulaic sequences and to make sure that the linguists in this field are talking about the same concept, Wray and Perkins (2000) define formulaic sequence as:

> *a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar* (Wray & Perkins, 2000, p.1).

In order to gain a comprehensive understanding of formulaic sequences, we will delve into their main characteristics. These characteristics encompass continuity, frequency, flexibility, semantic irregularity, and syntactic irregularity.

Continuity

As proposed by Wray and Perkins (2000) above, formulaic sequences can be continuous or discontinuous (Nattinger and DeCarrico, 1992; Wood, 2015; Wray & Perkins, 2000; Wray, 2002). It is usually more complex to identify and track discontinuous sequences than the continuous ones, as fillable slots and two-part expressions tend to blend into surrounding text - for example, 'not only ... but also' (Wood, 2015, p.9) and tracking discontinuous expressions require a more flexible approach to measurement (Wray, 2002, p.285). Another problem is that the importance of discontinuous sequences which consist predominantly or entirely of function words are usually overlooked, especially those 'logical connectors, subordinating and comparative units, and discourse markers that serve to connect longer stretches of discourse' (Yuldashev et al., 2013, p.42).

Frequency

Formulaic sequences are generally recurrent (Wray & Perkins, 2000). In corpus linguistics, by counting the frequency of co-occurrence between a target word and other words, in another word, the collocations, it can be seen that the collocations are not random and have a consistent frequency, indicating their common and idiomatic usage. Frequency plays a significant role in identifying formulaic sequences, as more frequent strings are likely to be stored and used prefabricated, suppressing alternative expressions.

Identifying formulaic sequences through computer searches is one of the most used method to detect the target sequences however such a method is not as straightforward as it may initially seem. Researchers must make decisions on what to include and set up the search accordingly, including determining the size of the word strings and the frequency threshold for inclusion. These frequency thresholds are somewhat arbitrary and depend on factors such as corpus size, desired data quantity, and the size of the chunks being searched. The length of recurrent word combinations tends to be inversely related to their frequency(Wray, 2002). Detecting discontinuous formulaic sequences in corpus can be even more tricky with the interruption of fillable slots. Another problem is that frequency is not the only factor relevant to capturing patterns of usage as many word strings are highly formulaic, but not frequent (e.g, long live the King) (Wray, 2002).

Flexibility

Some prefabricated strings are fully fixed in form, while others are freer, which are semi-preconstructed phrases that could be filled by morphological detail and/or open class items (Wray, 2002; Myles & Cordier, 2017). Howarth (1998) demonstrated an awareness of the continuum that exists between the fixedness and freeness of these lexical units, instead of a simple fixed or free two-way division. Table 1 shows that from left to right, the word combinations become more and more restricted. By using phrases derived from the same lexicon, such as 'blow' and 'under,' Howarth managed to show a clear and unified comparison of fixedness with examples.

Table 1. Collocational continuum (Howarth,1998, p.28)

	free combinations	**restricted collocations**	**figurative idioms**	**pure idioms**
lexical composites verb + noun	blow a trumpet	blow a fuse	blow your own trumpet	blow the gaff
grammatical composites preposition + noun	under the table	under attack	under the microscope	under the weather

In Wray (2002), flexibility refers to how formulaic sequences can be modified or adapted to fit different communicative situations or contexts. Some formulaic sequences are highly inflexible, meaning they can only be altered by losing their meaning or communicative function. For example, 'once upon a time' is a highly inflexible sequence that can only be easily modified by changing its meaning as a marker of the beginning of a fairy tale. Other formulaic sequences, however, are more flexible and can be adapted to fit different communicative contexts. For example, 'I'm sorry' can be modified by adding words such as 'really' or 'deeply' to express a greater level of apology. The more flexible formulaic sequences also include the ones with fixed frame and fillable elements, to be chosen according to the context needed, such as conjunctions. The most free formulaic sequences can be flexible in all parts but still with some limitation, such as the idiom principle (Sinclair, 1991)[1], and requires a relatively high frequency. Measuring the flexibility of formulaic sequences can be challenging, as it requires examining how well different variations of a given sequence are understood and used in different contexts.

Semantic irregularity

Similar as idioms and metaphors, some formulaic sequences have abandoned their semantic compositional meaning in preference for a holistic interpretation (Nattinger & DeCarrico, 1992, pp.32-33).This implies that when a formulaic sequence conveys a metaphorical meaning, comprehending it without ample context can pose a challenge for the listener or reader, which means that sequences that lack transparent meanings have to be idioms, else they would become unusable (Wray & Perkins, 2000).

Syntactic irregularity

Syntactic irregularity such as restriction on the normal scope for inflexional or transformational manipulation or a direct (object) appears with an intransitive verb, are are often observed in idiomatic expressions which aim to yield a holistic meaning (Wray & Perkins, 2000).

2.1.2 Categorisation of formulaic sequence

Given the intricate and extensive nature of formulaic sequences, achieving a comprehensive understanding requires a focused examination of specific types or categorizing them into groups with common characteristics. Consequently, while some investigations treat formulaic sequences as a holistic entity without detailed classification (Taguchi, 2008; Bardovi-Harlig, 2002), other studies concentrate on specific types, such as idioms, proverbs, collocations, and lexical phrases. Moreover, an increasing number of studies delve into examining their characteristics,

[1] Sinclair (1999) proposed the well-known idiom and the open choice principle, to investigate the role of collocation.The Open Choice Principle posits that language can be understood as the result of a series of complex choices.However, according to Sinclair, the open-choice principle cannot provide sufficient restraints on the selection of choices, another principle is necessary to be applied to explain the preference of using certain kind of string in texts by users, that is the idiom principle, which suggests that the phrases like ' of course' in which 'of' is not the preposition as applied in open-choice principle and 'course' is not a countable noun as dictionaries mention (pp110-111), should be considered as as a single word. The reason is that in 'of course' the elements have lost their semantic identity (pp110-111). Idioms, proverbs, clichés, technical terms, phrasal verbs ,collocations should all be treated likely.

encompassing form and/or function (Becker, 1975; Howarth, 1998; Sinclair, 1991; Nattinger & DeCarrico, 1992).

Classifications based on both form and function can encompass all the main characteristics of formulaic sequences. In terms of form, these sequences exhibit frequency, continuity, flexibility, semantic irregularity and syntactic irregularity. Concerning function, they involve frequency, continuity and semantic irregularity. For example, conjunctions, as identified by Nattinger and DeCarrico (1992), function as sentence builders and demonstrate both discontinuity and semi-fixed form, indicating characteristics of continuity and flexibility. On the other hand, the formulaic expression 'Once upon a time' is continuous, fixed in form, and conveys a holistic meaning distinct from the simple sum of its individual words, illustrating semantic irregularity.

One of the most significant studies on the classification of formulaic sequences is Becker's research in 1975. This study, although primarily emphasizes on lexical bundles, employs a form-function mixed approach, contributing significantly to our understanding of the categorization of formulaic sequences. It focuses on utterances which are formed through the repetition, modification, and concatenation of phrases consisting of multiple words that are already known. Noticing that much of our speech involves stitching together fragments of text that we have previously heard, while creative processes serve the purpose of adapting these existing phrases to fit new situations, Becker investigated the lexical phrases and proposed six major categories based on their natures and functions, in which nature represents the form.

Table 2. Six major categories of lexical phrases (Becker, 1975, p61)

Class	Nature (form)	Function	Examples
1.Polywords	Multi-word phrases admitting no variability, interchangeable with single words or concepts.	The same as single words.	- the oldest profession (n.) ='prostitution' - to blow up (vi., vt.) = 'to explode' - for good (adv.) = 'forever'
2.Phrasal constraints	Units consisting of a small number of words, some of which constrain the variability of others; in the limiting case the whole phrase is invariable.	Often specify how a particular expressive function is to be applied to particular semantic material.	-by sheer coincidence

3.Deictic Locutions	Phrases with low variability, short-to-medium length.	Serve as clauses or whole utterances whose purpose is to direct the course of conversation, i.e. the flow of expectations, emotions, attitudes, etc.	-for that matter = 'I just thought of a better way of making my point' - that's all = 'don't get flustered'
4.Sentence builders	Phrases up to sentence length, often containing slots for 'parameters' or 'arguments'.	Provide the skeleton for the expression of an entire idea.	(person A) gave (person B) a (long) song and dance about (a topic) = 'A tried to convince B of something, and was cynical and perhaps less than truthful about what he said'
5.Situational utterances	Usually complete little variability.	Utterances which are known to be the appropriate thing to say in certain circumstances; may be used out of context for effect.	how can I ever repay you? = expresses 'moderate-to-large gratitude in response to some kindness.'
6.Verbatim texts	Texts of any length memorized verbatim, or approximately so.	Used as substance for quotation, allusion, variation, and occasionally direct usage.	e.g.better late than never; How ya gonna keep 'em down on the farm?'

We can see from the above table that the Becker's categorisation describes details of each type by its nature (form) and function. The 'nature' gives restrictions of the fixedness, ranging from ploywords and situational utterances which permit no or little variability, deictic locutions are phrase with low variability, to the more free type sentence builders which allows slots inside (although not being pointed out clearly by Becker, just as another category verbatim texts). Phrasal constraints, the variability varies internally and in some cases it is invariable. The categorization based on 'nature' also provides insights into the length of words within each category. Polywords, for instance, can be some short multi-word units . Deictic locutions, on the other hand, tend to be relatively short to medium in length. Sentence builders, as their name suggests, have the capacity to form phrases that are sentence-length. Furthermore, verbatim texts encompass texts of any length, allowing for a broader range of linguistic expression.

In Becker's study, the inclusion of various functions provides valuable assistance in classifying and identifying lexical phrases. However, it is important to note that these functions encompass a blend of grammatical, semantic, and pragmatic aspects, which can sometimes make the categorization less transparent and straightforward.

One of the important of few textbooks about lexical phrase is Nattinger & DeCarrico's study in 1992, in which they also investigate the lexical phrases from both their form and function.

To classify by form, Nattinger & DeCarrico firstly distinct lexical phrase from syntactic strings (e.g. NP+Aux +VP) and collocations without particular pragmatic functions(e.g. rancid butter) and define lexical phrase as a type of collocation with assigned pragmatic functions. Four structural criteria were employed to classify the lexical phrases into four categories- polywords, institutionalized expressions, phrasal constraints, and sentence builders. These criteria encompassed an examination of the phrases' length and grammatical status, a determination of whether they possessed a canonical or non-canonical shape, an assessment of their variability (variable or fixed), and an evaluation of their continuity (continuous or non-continuous), polywords , phrasal constraints, and sentence builders.

(1) **Polywords**:fixed and continuous short phrases which function as individual lexical items. They are both canonical and non-canonical. For example: by the way, hold your horses (canonical); as it were, so far so good (non-canonical). All linking devices(or 'relators') such as *however,moreover, nevertheless* are considered as polywords.

(2) **Institutionalized expressions**: sentence length invariable phrases which are mostly canonical and continuous, functioning as separate utterance.

This category includes proverbs, aphorisms and formulas for social interactions, as well as chunks which are stored as units, such as *how do you do? have a nice day, be that as it may, long time no see.*

(3) **Phrasal constraints**: short-to-medium-length phrases which allow variation of lexical and phrasal categories. Phrasal constraints can be both canonical and non-canonical. One example of canonical phrasal constraint is *a___ago* (canonical)*,*in which the slot can be filled by *day, year,* or *a very long time*, etc. Another example is non-canonical: the___er the ____er, in which the phrase can be constructed to *the sooner the better*, *the busier the happier*, etc.

(4) Sentence builders: frameworks for the whole sentences which allow considerable variations such as phrasal (NP,VP) and clausal(S) elements. The sentence builders can be both canonical and non-canonical with continuous or discontinuous pattern.

Compared to Becker, Nattinger & DeCarrico proposed a clear linear of variation and discontinuity of the four categories, that is, an increasing possibility of variation and discontinuity. However, as noted by Nattinger and DeCarrico, the establishment of clear boundaries between different categories proves challenging due to inherent overlap and a lack of clarity. Despite applying the mentioned criteria, the delineation of boundaries between categories remains somewhat elusive.

To integrate the classification of form and function, Nattinger and DeCarrico (1992) identified three distinct functional categories which are social interactions, necessary topics and discourse devices. Further more, they explained how the three forms polywords (institutional expressions are considered as sentence length ploywords), phrasal constraints and sentence builders are presented in each category, as shown in the following Table 3.

Table 3. Categorisation of lexical phrases by form & function (Nattinger & DeCarrico, 1992)

form function	social interactions	necessary topics	discourse devices
polywords	by the way (shiting a topic) all right? (checking comprehension)	a great deal (quantity) not too expensive (shopping)	in other words (exemplifier) at any rate (fluency device)
phrasal constraints	_____me?(clarifying:audience) see you ____(parting)	I'm from____ (autobiography) how much is ____(quantity)	as far as I _____(evaluator) as a result of ____(logical connector)
sentence builders	what I mean is X (clarifying: speaker) do you know X? (nominating a topic)	what do you like to X? (likes) what time X? (time)	there's no doubt that X (evaluator) my point here is X (summarizer)

2.1.3 Identification and classification of formulaic sequence in Chinese

Identification

In Chinese, the term most commonly used to indicate formulaic sequences is 'yǔkuài' (语块) or 'yùzhì yǔkuài' (预制语块) (Wang, 2017; Qian, 2008; Zhou, 2007). Research on formulaic sequences in Chinese has been relatively limited compared to that conducted on English and other languages. There have been fewer textbooks dedicated to Chinese formulaic sequences, as, for a long time, the concept of *yǔkuài* was blended into duǎnyǔ' '短语'(short phrase) in the study of the Chinese language lexicon. It was only when the learning of Chinese as a second language began to receive more attention that the concept of *yǔkuài* emerged, primarily based on the terminology and identification of formulaic sequences in English. Consequently, these studies primarily concentrate on L2 teaching rather than L2 learning. Nonetheless, valuable insights can still be derived from the research.

Yǔkuài, similar to formulaic sequences in English, lacks a unified and widely accepted identification in Chinese. Zhou (2007, p.100) pointed out that *yǔkuài* is different from *cízǔ* or *duǎnyǔ* in traditional grammar, but he did not give further explanation on the detailed difference. According to Zhou, *yǔkuài* includes the semantic units or the fixed structures. Since in second language teaching, we usually focus on the form or structure of these units , Zhou gives the definition of *yǔkuài* as units which are larger than individual words - 'cí' (词) and frequently appear in various types of sentences and serve as building blocks of sentence structure.

However, this definition suffers from certain shortcomings and vagueness. For instance, as highlighted by Wray (2002) and Wray & Perkins (2000), relying solely on frequency as a criterion to define formulaic sequences can be limiting. Some highly formulaic prefabricated strings may not occur frequently in natural language corpora, thus posing a problem when using the term 'frequently' to define *yǔkuài* (Wang, 2017).

Similarly, Qian (2008) approached the definition of *yǔkuài* from the perspective of L2 teaching. Qian provided a list of language structures that can be considered as *yǔkuài*, including collocations, idioms, proverbs, maxims, catchphrases, children's songs, song lyrics, religious scriptures, and more. Building upon Wray's definition, she considered these *yǔkuài* as language structures composed of multiple words, stored, extracted, and used as a whole, each occupying different grammatical levels. Notably, Qian's definition encompasses language structures of discourse length, such as children's songs, song lyrics, and religious texts, as part of the *yǔkuài* category. However, this broad definition of *yǔkuài* risks of over-inclusion. Moreover, unlike Wray's definition, which is grounded in the characteristics of formulaic sequences, Qian's definition of *yǔkuài* is descriptive in nature and relies on listing *yǔkuài* with different length.

Among the various studies conducted on this subject, one notable example is the comprehensive investigation carried out by Wang (2017). In his research, Wang defined *yǔkuài* as 'a prefabricated non-word sequence consisting of two or more morphemes, which is no longer than a sentence, is often stored and retrieved as a whole during usage.'

In order to enhance our comprehension of formulaic sequences in the Chinese language and evaluate the precision of Wang's definition, we need to clarify the components that can be confused with yǔkuài in Chinese. These components include *zì* '字' (character or morpheme), *cí* '词' (word), *duǎnyǔ* '短语' (short phrase), *chéngyǔ'*

成语' (idiom), *yànyǔ* '谚语' (proverb) and *jùzi* '句子' (sentence), as well as their relationships with yǔkuài in Chinese.

***Yǔkuài* VS *zì* (character or morpheme)**

In Old Chinese, the equation 'one character = one syllable = one word' is commonly applicable (Shi, 2002). However, in modern Chinese vocabulary, the majority of words consist of disyllabic (and occasionally trisyllabic) structures (Zhongguo Wenzi Gaige Weiyuanhui Yanjiu Tuiguang Chu, 1959). Therefore, while one character still corresponds to one syllable, it does not always correspond to one word, especially in compound words composed of two morphemes or more (Kecskes & Sun, 2017).

As for morphemes, one character does not necessarily equal one morpheme. For instance, words like *húdié* '蝴蝶'(butterfly) and *biānfú* '蝙蝠'(bat) are multisyllabic and, thus, multi-character words but consist of a single morpheme because *hú* and *dié* can only be used together as a unit. This example indicates that for more precision, it is preferable to consider 'morpheme' as the fundamental element when defining the boundaries of *yǔkuài*. This leads to a question: Can a single morpheme constitute a *yǔkuài*, especially when the morpheme is composed of more than one character? The answer, as outlined in Wang's research, is unequivocally negative, as he stipulates the requirement of 'two or more morphemes' instead of 'two characters' for the formation of a *yǔkuài*.

***Yǔkuài* VS *cí* (word)**

According to Zhang et al. (2000), a word is composed of morphemes and represents a unit syntactically higher than a morpheme. A word represents a specific semantic content, has a fixed phonetic form, and is the smallest linguistic unit that can be used independently. When a morpheme is used independently, it becomes a word, for example, *tiān*'天' (sky). However, when a morpheme is not used independently but is combined with other morphemes to form a word, it functions solely as a morpheme. For instance, in the word *tiānkōng* '天空' (sky), *tiān* is just a morpheme instead of a word.

Can a word be considered as *yǔkuài*? To address this question, we must examine the various forms of words:

- A word composed of a single morpheme cannot be categorised as *yǔkuài*.

- A word composed of two or more morphemes has the potential to be classified as *yǔkuài*.

We observe that ,syntactically, from the word level, it starts to show a potential overlap between the boundaries of words and *yǔkuài*. This indicates that there exists a possibility for the distinction between words and *yǔkuài* to become blurred. However, Wang (2017) posits that when considering a word composed of multiple morphemes, the primary classification should be as a word rather than as a *yǔkuài*.

***Yǔkuài* VS *duǎnyǔ* (short phrase)**

From the perspective of form, both *yǔkuài* and phrases are comprised of multi-morphemic combinations known as 'non-words'. The similarities between *yǔkuài* and short phrase contribute to the difficulty of distinguishing them from each other. The overlapping nature of these linguistic units arises from their shared characteristic of being composed of multiple morphemes. However, there are still some distinctions between them, as indicated by Wang (2017) in the following.

Structurally, a short phrase represents a direct and typically continuous combination of words. In contrast, yǔkuài, such as *jù...shuō* '据……说' (according to...) and *yuè...yuè...*'越……越……' (the more...the more...), can also exhibit discontinuity. This viewpoint aligns with Wray's (2000) definition of formulaic sequences, which encompasses both continuous and discontinuous sequences.

Semantically, the meaning of a short phrase usually closely corresponds to the meaning of its constituent parts. However, yǔkuài tends to function as an indivisible whole.

In terms of frequency, *yǔkuài and* short phrase share more in common. When a short phrase is frequently used, its cohesion strengthens along with the increase in usage frequency. Consequently, the elements within the short phrase become more fixed, leading to the possibility of certain highly frequent collocations with a high co-occurrence rate being considered as *yǔkuài*.

***Yǔkuài* VS chéngyǔ(idiom)**

According to Wang's (2017) definition of formulaic sequences, we can observe that idioms can be considered one type of yǔkuài, such as the four-character idiom *rén shān rén hǎi* (人山人海), which means 'crowded with people'. These are 'prefabricated non-word sequences consisting of two or more morphemes, ' utilized as complete units. However, the scope of *yǔkuài* is much wider, encompassing other units such as *zì* '字' (character or morpheme), *cí* '词' (word), *duǎnyǔ* '短语' (short phrase), *chéngyǔ'* 成语' (idiom), *yànyǔ* '谚语' (proverb) and *jùzi* '句子' (sentence), as well as their relationships with yǔkuài in Chinese.

***Yǔkuài* VS yànyǔp (proverb)**

Similar to idioms, proverbs are one type of yǔkuài, and yǔkuài contains more concepts than proverbs. It needs to be noted that a proverb sometimes can appear as a sentence, such as *饭后百步走, 活到九十九* (Take a hundred steps after a meal and live to ninety-nine), which means 'taking a walk after meals may lead to a longer life', while *yǔkuài* usually consists of sequences shorter than a sentence length. The reason why this is not a problem is that proverbs can only be used as a whole unit to convey a fixed meaning, without any modification. Therefore, proverbs are different from normal sentences.

***Yǔkuài* VS jùzi (sentence)**

A sentence is the fundamental unit of language communication. In terms of its form, a sentence has a unified intonation that represents a certain mood. In written form, this is reflected by the use of punctuation marks at the end of a sentence, such as a period (.), question mark (?), or exclamation mark (!). A sentence is capable of expressing a relatively complete meaning and fulfilling a simple communicative task (Zhang et al., 2000, p.30). According to Wang, *yǔkuài* are typically smaller than sentences, but they can still appear as complete sentences without surpassing the sentence level.

In summary, formulaic sequence in Chinese (which is yǔkuài) can be defined as 'a prefabricated non-word sequence consisting of two or more morphemes, which is no longer than a sentence, is often stored and retrieved as a whole during usage.' Yǔkuài can include zì '字' (character or morpheme), cí '词' (word), duǎnyǔ '短语' (short phrase), chéngyǔ' 成语' (idiom), yànyǔ '谚语' (proverb) and jùzi '句子' (sentence) when these components contain more than one morphemes and are used as a whole unit.

Classification of formulaic sequence in Chinese

As discussed in 2.1.2, formulaic sequences, as an umbrella term, are typically categorised into different groups for further investigation based primarily on their forms and functions. Likewise, *yŭkuài* in Chinese can be also classified into comparable sub-categories utilizing a similar classification system.

The study of Wang (2017) offers a comprehensive description of how *yŭkuài* are divided into the following subcategories based on their form and function. Wang's research provides valuable insights into the systematic categorization of *yŭkuài*, contributing to a deeper understanding of their linguistic characteristics and usage patterns.

(1) *dāpèi* '搭配'(collocation):

- *dǎzhāohū* '打招呼' (lit. beat greet): greeting someone.
- *yǒuyìsi* '有意思' (lit. have interest): interesting; meaningful.
- *nào xiàohuà* '闹笑话' (lit. create jokes): make a fool of oneself.
- *chī yādàn* '吃鸭蛋' (lit. eat duck eggs): get zero score on an exam.

***(2) kuàngjià géshì* '框架格式' (frame structures):**

The framework patterns can be broadly classified into two categories: phrase framework and sentence framework. The phrase framework pertains to the structural organization at the phrase level, such as 'Verb + what' (e.g., *eat what* means 'what to eat'). The sentence framework encompasses cohesive structures and special sentence patterns, such as 'if...then...' and 'not only...but also...' Typically, the framework format comprises fixed elements and non-fixed elements (or 'filler items'). The fixed elements cannot be replaced or exhibit poor substitutability, while the content in the non-fixed elements is subject to varying degrees of constraints, displaying close relationships in terms of structure and possessing a fixed framework meaning.

(3) *xíguànyòngyǔ* '习惯用语' (conventionalized expressions):

Conventionalized expressions , such as *méitóuméinǎo* '没头没脑' (without head or tail, which means 'something that lacks structure, coherence, or logical order') and *méiwánméile* '没完没了' (endlessly), fall under this category.

(4) *shúyǔ* '熟语' (idioms):

Familiar expressions in Chinese generally encompass 4-character idioms, such as *tiānhándì dòng* (lit. sky cold, ground frozen), which is used to describe extremely cold weather. Other types of idioms include proverbs, riddles, maxims, famous quotations, aphorisms, as well as signage expressions like 'No Smoking'.

(5) *tàoyǔ* '套语'(polite formulas):

Polite formulas, refer to single fixed phrases, including greetings and blessings. Examples include:

zhù nǐ yīlù shùnfēng! '祝你一路顺风!' (Wishing you a smooth journey!)

xīnnián kuàilè! '新年快乐!' (Happy New Year!)

hòu huì yǒu qī! '后会有期!' (See you later!)

- nǐ hǎo! - nǐ hǎo! - '你好!' - '你好!' (- Hello! - Hello!)

yī duì bùqǐ! yī méi guānxi! - '对不起!' - '没关系!') (- Sorry! - It's okay!)

Similarly, through a comprehensive analysis of the form and functions of *yǔkuài*, Wang (2020) has skilfully categorized them into distinct groups (pp. 46-48). Wang's research stands as one of the most reputable and dependable studies concerning *yǔkuài*, providing valuable insights into their structural and functional aspects.

These categories are:

(1) *shúyǔ* '熟语'(idioms), including three subcategories:

(i) *chéngyǔ* 'idioms which are normally four characters' (成语): such as fēngkǒu làngjiān'风口浪尖 (lit. where the wind and the waves are the fiercest) which means 'at the heart of the struggle'.

(ii) *guànyòngyǔ* '惯用语' (institutionalised expressions):

such as 'pèng dīngzi' '(lit. bump one's head against a nail), receive serious rebuff'.

(iii) qítā lèixíng shúyǔ '其他类型熟语'(other types of idioms): such as tiān shàng diào xiànbǐng '天上掉馅儿饼' (lit. falling pie from the sky) which means 'good things come without any reasons, or things with a slim chance which hoped by fantastic people'.

(2) tàoyǔ '套语' (polite formulas): *huānyíngguānglín* '欢迎光临'(welcome); *máfannín* '麻烦您' (sorry to have troubled you).

(3) chārùyǔ '插入语' (parentheses): *huànyánzhī* '换言之'(in other words); *jùwǒ suǒzhī* '据我所知'(as far as I know).

(4) conventionalized expressions (习用语 xíyòngyǔ):*bùyàyú* '不亚于'(nothing less than); *cóngbù* '从不'(never).

(5) *gāopín dāpèi* '高频搭配' (high-frequency collocations) :*chōngmǎn huólì* '充满活力'(full of energy); *dìngyuè zázhì* '订阅杂志' (subscribe to a magazine); *gǔqǐ yǒngqì* '鼓起勇气' (gather courage).

(6) *kuàngjià jiégòu* '框架结构' (frame structures) with two subcategories

(i) duǎnyǔ kuàngjià '短语框架' (phrase frames): *bāokuò......zàinèi* '包括...在内'(including...inside); *gēn......yīyàng* '跟..一样'(the same as...).

(ii) sìzìgé '四字格' (four character frames): qī ... bā ... '七...八... ' (seven... eight... functions as a frame in qīshàngbāxià '七上八下' (lit. seven up eight down), which means 'nervous and uneasy'.

(7) *liàngcí duǎnyǔ* '量词短语' (classifier phrases) with two subcategories:

(i)Nominal classifier phrases :yī chǎng xuě '一场雪' (a fall of snow).

(ii)Verbal classifier phrases :bái yī yǎn '白一眼' (roll eyes at somebody).

2.1.4 Fixed, semi-fixed and free form

Building upon the significant contributions of Wang (2017) and Wang (2020) in the classification of formulaic sequences based on form-function, this study proposes a classification considering both form and function as well. Given the extensive range of formulaic sequences, our intention is to classify and select the most representative formulaic sequence from each sub-category. This approach ensures that the chosen sequences not only embody the characteristics of their respective groups but are also easily extractable from the corpus for thorough examination..

One important feature of formulaic sequences which shows in its form but linked to its function is the flexibility. Summarising the sub-categories identified in Wang (2017) and Wang (2020), it can be seen that they share some features in common according to the flexibility, as shown in Table 4. In this table, polite formulas, idioms, conventionalized expressions and parentheses are indicated as formulaic sequences with fixed forms by studies, sentence frames are semi-fixed sequences with the fixed frame and fillable slots, high frequency collocations are the most free form which requires a certain level of frequency.

Table 4. Subcategories of *yǔkuài* according to their fixedness

Subcategories	Wang (2017) ; Wang (2020)	Wang (2017)	Wang (2020)
fixed form	polite formulas [2] idioms[5] conventionalized expressions [6] Parentheses[7]	collocation[3]	classifier[4]
semi-fixed form	frame structures[8] (the frame itself is fixed but the fillable slots are flexible)	collocation[9]	
free form	high frequency collocation[10]		

In terms of the fixed form, one of the most typical examples of formulaic sequences, referred to as *yǔkuài* in Chinese, is seen in the four-character idioms, known as *sìzìchéngyǔ* '四字成语'. This is because *sìzìchéngyǔ* represents a traditional expression in the Chinese language, marked by its distinct structure of exactly four characters. These idioms maintain a fixed form and find their roots in classical Chinese literature, drawing inspiration from influential sources associated with Confucianism, Daoism, and Buddhism. The substantial impact of these three major philosophical and religious traditions on the cultural landscape of East Asia is evident in the broad range of cultural references embodied by *sìzìchéngyǔ* idioms. This reflects the abundant heritage and profound philosophical insights embedded in traditional Chinese literature and thought (Conti, 2017).

Regarding the semi-fixed form, conjunctions not only display a certain degree of flexibility but also convey the property of continuity, which can take both continuous and discontinuous forms. While conjunctions maintain a fixed nature as connectors of phrases or clauses, the elements they link can vary as needed.

[2] Examples of polite formulas: huānyíng guānglín '欢迎光临' (welcome); lù shàng xīnkǔ le '路上辛苦了' (you must have had a tough journey).

[3] Examples of fixed form collocations: dèng yǎnjīng '瞪眼睛 ' (stare eyes) in which *dèng* (stare) can not be substitued by other verbs to maitain the same meaning.

[4] Examples of classifier: yī chǎng xuě '一场雪' (a fall of snow' ;yī fù yǎnjìng'一副眼镜' (a pair of glasses).

[5] Examples of idiom: táolǐ mǎn tiānxià '桃李满天下' (lit. peach trees and plum trees are everywhere) , which means 'having students all over the world' ;báirì mèng '白日梦' (daydream); céngchū bùqióng '层出不穷' (emerge in an endless stream),

[6] Examples of conventionalized expressions: bùyòng shuō '不用说' (let alone); bùwàihu '不外乎' (nothing more than)

[7] Examples of parentheses: huànyánzhī '换言之' (in other words); jíbiàn rúcǐ 即便如此 (in spite of that).

[8] Examples of frame structures: bāokuò……zàinèi '包括......在内' (including……'; yǐ……ér gàozhōng '以......而告终' (end up with……).

[9] Examples of semi-fixed form collocation: tóuzī gùwèn '投资顾问' (investment adviser)', tóuzī '投资' (investment' can be substituted by jīngjì '经济' (economics), fǎlǜ '法律' (law), ānquán'安全' (safety').

[10] Examples of high frequency collocation:chōngmǎn huólì '充满活力' (full of energy); gǎn shímáo '赶时髦' (keep up with fashion) ;chuántǒng wénhuà '传统文化' ('tradition culture).

In the context of the free form, Wang (2017) and Wang (2020) propose a classification that designates only high-frequency collocations to fall within this category. However, the range is still too vast for measuring and analyzing formulaic sequences. Therefore, we narrow down our research to examine verbs and their collocations with high frequency because a verb is mandatory in almost every sentence. Moreover, we specify our selection to V+N$^{(object)}$ high-frequency collocations since the N$^{(object)}$, as the receiver or patient of the action, shows a tight relationship to its verb.

Consequently, in this study, we classify the formulaic sequences in Chinese into three types: fixed, semi-fixed, and free form. In each of these types, we choose four-character idioms, conjunctions, and V+N$^{(object)}$ high-frequency collocations as the most featured examples.

2.3 L2 Formulaic sequence learning in SA context and its influence on writing skills

Having gained an understanding of formulaic sequences and their categorization based on distinct characteristics, the subsequent section revolves around elucidating the significance of formulaic sequences in the context of second language learning in study abroad settings. It focuses on two dimensions: the significance and challenges of learning Chinese formulaic sequences in an SA context, and the influence of formulaic sequences on L2 writing skills. The aim is to contribute insights into the unique landscape of L2 Chinese formulaic sequence learning during study abroad experiences, shedding light on its impact on writing proficiency.

2.3.2 L2 Chinese formulaic sequence learning in SA context

i. L2 learning in SA context

Study abroad refers to 'a temporary sojourn of pre-defined duration, undertaken for educational purposes' (Cordier, 2013, p.11). Study Abroad can offer a valuable (immersion) environment, seen as beneficial for language development (Freed, 1995; Freed et al,. 2004; Kinginger, 2008; Kinginger, 2011;Wright & Schartner, 2013) , and specifically for communicative competence (Celce-Murcia, Dornyei, & Thurell, 1995; Kinginger, 2011, 2018; Wright, 2019, 2020). Communicative competence, encompassing the ability to express, interpret, and negotiate meaning (Savignon, 2017), is often defined with reference to appropriate language use. This entails the skill of matching language forms to social contexts without unintentionally violating expectations (Kinginger, 2009). Experiences of study abroad go beyond the typical classroom language use, expanding and refining students 'linguistic repertoires' (Kinginger, 2009). While study abroad participants may not perform speech acts like native speakers, their exposure to diverse practices and the observation of deviations from their expectations can enhance their communicative flexibility (Kinginger, 2009).

Among the benefits that study abroad can bring to L2 learning, most of those mentioned above primarily emphasize improvements in speaking. Improvement in writing proficiency has not gained as much attention as speaking. In our study, we aim to investigate the development of lexical and syntactic complexity in writing within a study abroad setting, in comparison to improvements in speaking, commonly assessed through measures of fluency.

The role of study abroad experiences in facilitating second language (L2) learning has been extensively discussed in the context of English and other Indo-European languages, such as French and Spanish. Similarly, in the case of L2 Chinese learning, there is evidence suggesting that study abroad in China can contribute to language development. For instance, Du (2013) conducted a study in which he observed that an increased amount of time spent speaking Chinese, both within and outside of formal classroom settings, correlated with enhanced fluency among participants. The findings of this study underscore the significance of 'time-on-task' as a crucial factor influencing fluency development. Additionally, further research, as demonstrated by Wright (2019), has explored task effects and their positive impact on learners oral proficiency during their study abroad experiences.

ii. The role of formulaic sequence in SLA

In recent years, an increasing number of studies have made use of corpus data to add weight to the importance of multi-word units in language. Research shows that more than one third of language phrases used in speech are formulaic (Erman & Warren, 2000; Howarth,1998).

Formulaic sequences are also essential in language learning (Ellis, 1996 and 2012). A significant part of language learning involves memorizing sequences of language, including vocabulary (such as phonological units and their sequences) and discourse (such as lexical units and their sequences within clauses and collocations) (Ellis, 1996; Sinclair, 1996). Moreover, the memorized short-term gains can turn into long-term knowledge and contribute to more freely creative language use. Such benefits can be found in both L2 speaking and writing learning.

Myles et al. (1999) conducted a study on the significance of chunks, which represent one type of formulaic sequence, in second language learning. They investigated the relationship between chunks and the creative construction of language in second language acquisition. Creative construction refers to the L2 learner's ability not only to learn formulaic chunks but also to break them down and reconstruct them, or to employ the chunks in a more flexible manner based on the contextual requirements. This suggests a higher level of proficiency in mastering the target language. To investigate how chunks relate to the creative construction , Myles et al., analysed the use of interrogatives in a large collection of L2 French utterances produced by learners in early classroom settings. The results show a strong relationship between creative construction of language and the breakdown of chunks, in which chunks serve as a database for the more complex creative utterances. The interrogative chunk ,as an example, which plays a fundamental role in the subsequent analysis and creative production of language.

For instance, learners initially use more complex interrogative structures, such as wh-fronted interrogatives with inversion, without fully analysing them. However, as their language proficiency improves, they gradually modify these formulaic structures to align with the basic canonical order of the language. The learners do not completely abandon these structures but instead continue to work on them, making certain modifications to enhance their creative competence in language construction.

Such a process is a gradual one as the learners will modify the patterns when they when find it inappropriate according to the references. Those learners who were able to successfully memorize formulas and continued to work on them throughout the study were also the ones who demonstrated early engagement in creative construction and made significant progress along the developmental continuum over the two-year period.

Ellis (1996) also found the crucial role of learning language sequences in language acquisition, serving as the database, that is considering chunking as a general process in second language acquisition and by internalizing and utilizing these chunks of language, learners' long-term knowledge of lexical sequences in formulaic phrases can develop a solid database that contribute to their acquisition of language grammar (Ellis, 2012).

iii. L2 Influence of SA on FS improvement in writing

Several studies explore L2 formulaic sequence knowledge with a focus on L2 written production, emphasizing the importance of study abroad (SA) influence on formulaic sequence development in writing. Although these studies only investigate one specific area of formulaic sequences, they still shed light on the role SA can play in improving the use of formulaic sequences in writing. As mentioned in Siyanova-Chanturia's study, Waibel (2008) investigated the use of phrasal verbs (one type of formulaic sequence) in L2 writing and found that immersion in the target-language country facilitates the acquisition and incorporation of phrasal verbs into written communication. Another similar study by Groom (2009) examined Swedish learners of English, dividing them into two groups based on their exposure to English-speaking environments: one group with less than one month and another with 12 months or more. Through statistical analyses, Groom found a positive correlation between time spent in an English-speaking environment and collocational accuracy in the learners' writing. Specifically, the group with over 12 months of exposure demonstrated higher collocational accuracy compared to the group with less than one month, indicating that extended exposure enhances the natural and accurate use of word combinations in writing (Siyanova-Chanturia, 2015).

iv. L2 Chinese formulaic learning in SA context

Formulaic language has gained considerable empirical interest recently, primarily because it is recognized as a crucial component of L2 learners 'communicative competence' (Schmitt&, 2004; Wray, 2002) as indicated by Taguchi, Li, and Xiao in 2013.

A study conducted by Taguchi, Li, and Xiao (2013) explored the development of L2 Chinese formulaic competence within the context of study abroad. The research involved 31 American students studying Chinese at an intermediate level in a university in China. During their semester-long study abroad, the participants completed a computerized speaking test consisting of 24 formulae-use situations, wherein they produced formulaic expressions. Native speakers then evaluated the appropriateness of their production on a four-point scale, and planning time was also measured. Additionally, a survey was administered to assess the learners' perceived frequency of encountering formulae-use situations.

The results of the study showed significant gains in appropriateness and fluency of formulae production among the learners. Interestingly, the reported frequency of encountering target formulae-use situations did not correlate with the observed gains, except for learners with lower pretest scores. A qualitative analysis of the data revealed four patterns of change: (1) progress towards target formulae, (2) progress towards target-like slot-and-frame patterns, (3) progress towards non-target formulae, and (4) stabilized non-target formulae use.

The investigation contributes valuable insights into the development of formulaic competence in L2 Chinese learners during study abroad experiences. The findings shed light on the factors influencing the acquisition of formulaic language and provide a deeper understanding of learners' formulaic language production within the study abroad context.

Taguchi (2011), investigated the effect of proficiency and study abroad experience on the comprehension of routines (formulaic expressions) and implicature (non-formulaic, non-literal utterances) among Japanese college students studying English as a foreign language (EFL). Participants were divided into three groups based on their proficiency levels and study abroad experiences. Results indicated that study abroad experience had no effect on response times in listening tests but did significantly influence the accurate comprehension of routines. However, the indirect examination of study abroad effects due to participants' returnee status limited the study's findings, highlighting the need for studies recruiting participants while abroad to establish a more direct relationship between residence abroad and formulaic competence.

Additionally, Dornyei et al. (2004) provided empirical evidence supporting the notion that increased target language input, acquired through participation in real-life communicative events, facilitates the transition from idiosyncratic, inter-language usage of formulae to more conventional, target-like usage. However, the relationship between the development of formulaic competence and the amount of L2 exposure remains underexplored, warranting further research in this area.

Taguchi (2011)'s study aimed to investigate the development of L2 Chinese formulaic competence during study abroad. They selected candidate formulae commonly used in daily situations in China and created scenarios to elicit formulaic expressions. Results revealed significant gains in appropriateness scores and frequency of target formulae among learners during the study abroad period. Moreover, the learners exhibited increased planning speed, indicating improved retrieval of lexico-syntactic knowledge necessary for target formulaic expressions. These findings align with prior research highlighting the role of study abroad contexts in promoting L2 formulaic competence. Regarding the relationship between the perceived frequency of encountering target formulae-use situations and gains in production, Pearson correlation analysis applied in Taguchi's study did not show significant correlations. However, post hoc regression analysis revealed that learners' initial scores and perceived frequency of encounter together explained the gains. This suggests that learners' initial level of formulaic competence in relation to the frequency of encountering target formulae-use situations likely influenced the observed gains. Qualitative analyses further revealed other influential factors, such as lexical and syntactic knowledge, as well as knowledge of form-context mappings. These findings underscore the multifaceted nature of formulaic competence development in L2 learners during study abroad experiences.

v. Challenges of formulaic sequences learning

Despite the importance of formulaic sequences in writing, L2 learners often encounter challenges when it comes to effectively utilizing these sequences (Li & Schmitt, 2009). For example, Li and Schmitt (2009) pointed out that 'learning to write well also entails learning to use formulaic sequences appropriately,' and 'L2 learners' failure to use native-like formulaic sequences is one factor in making their writing feel nonnative' (p. 86). Another problem is the lack of diversity when using those sequences (De Cock et al., 1998) and overuse some particular ones which they are familiar with (Granger, 1998; Li & Schmitt, 2009).

Moreover, various types of formulaic sequences can serve distinct functions in writing performance. Fixed formulaic sequence, like Chinese four-character idioms, act as 'building blocks' (Arnon & Christiansen, 2017) conveying complete meanings. Semi-fixed formulaic sequence, such as conjunctions, offer flexible sentence structures, and more flexible formulaic sequence, like V+N$^{(object)}$ high frequency collocations, allow creativity with constraints.

L1 transfer

One of the most discussed factors contributing to the challenges encountered by L2 Chinese learners in acquiring formulaic sequences is L1 transfer. The study of He (2023) undertakes a comprehensive exploration of the challenges encountered by second language learners in acquiring Chinese directional complement (DC) constructions, a subset of multiword sequences (formulaic sequences). These constructions, resembling single words but conveying intricate literal and figurative meanings, pose difficulties for L2 learners. Employing a situation-cued sentence completion test and a grammaticality judgment test, the analysis of 58 L1 English learners of Chinese revealed that the learners' acquisition of DCs was impacted by both the syntactic complexity of DC constructions, and the learners' L1 experience.

The study shows that English learners of Chinese experienced challenges in comprehending and producing DC constructions influenced by their L1 experience. Specifically, the expression of the function of 'lái' [hither] and 'qù' [thither] in English, particularly through the one-to-one form-function mapping of 'come'and 'go,' hindered the learners' understanding of these elements when functioning as DCs. Additionally, the presence of these elements in compound DC constructions with noun phrases (NPs) was overshadowed or obstructed by cues more familiar to learners in their L1, such as discourse context, where their function is reliably conveyed through prepositional phrases in English.

2.3.3 Formulaic sequence and L2 writing skills

i. The role of formulaic sequence on L2 writing development

'Knowledge of vocabulary is obviously a prerequisite for writing' (Li & Schmitt, 2009, p.85) which often encompasses formulaic multi-word sequences (Nattinger & DeCarrico, 1992; Sinclair, 1991; Wray, 2002). Corpus linguistics shows abundant occurrence and functional significance of lexical phrases in written text (Erman & Warren, 2000). According to the research conducted by Erman and Warren (2000), it was discovered that more than half of the written discourse analyzed consisted of formulaic sequences, therefore it is imperative to explore their role in the development of second language writing (Li & Schmitt, 2009). Coxhead and Byrd

(2007, pp. 134–135) pointed out that the significance of word sets (formulaic sequences) for L2 learners students are due to several reasons:

> *(a) the word sets are often repeated and become a part of the structural material used by advanced writers,making the students' task easier because they work with ready-made sets of words rather than having to create each sentence word by word;*
>
> *(b) as a result of their frequent use, such [sequences] become defining markers of fluent writing and are important for the development of writing that fits the expectations of readers in academia;*
>
> (c) *these [sequences] often lie at the boundary between grammar and vocabulary; they are the lexicogrammatical underpinnings of a language so often revealed in corpus studies but much harder to see through analysis of individual texts or from a linguistic point of view that does not study language-in-use. (Coxhead & Byrd, 2007, pp. 134–135)*

Khoualdi pointed out that formulaic sequences can not only be a sign of fluent writing that meets the expectations of readers in academia (Coxhead & Byrd, 2007) but also aid in achieving accuracy in writing. Since formulaic sequences are retrieved as wholes from memory, they are more likely to reduce grammar errors and odd word combinations (Khoualdi, 2017).

El-Dakhs (2017) investigated the effect of explicit instruction of formulaic sequences in pre-writing vocabulary activities on L2 writing and found that formulaic sequences can have a positive influence on the learners’ lexical choices and overall writing quality (skills).

ii. Measurement of L2 writing skills

The influence of SA on L2 writing skills can be measured by lexical complexity, fluency, accuracy.

Lexical complexity

The increasing recognition of vocabulary's significant role in learners' perceptions has been a primary factor driving its focus. In the study conducted by Ife, Vives Boix, and Meara (2000), it is examined how vocabulary development and language proficiency progress in the study abroad context. They aimed to assess the rate and type of progress made by learners at different proficiency levels. The findings revealed that study abroad participants tend to develop more extensive lexical repertoires compared to learners who receive solely classroom instruction. Additionally, these participants' lexicons showed a tendency to be organized in a manner that resembles native speakers.

Ife, Vives Boix, and Meara utilized the Eurocentres Vocabulary Size Test to estimate the vocabulary size of 53 European exchange participants studying in Britain. The results demonstrated a remarkable improvement in scores for the entire group, with some students more than doubling their initial vocabularies. On average, the rate of vocabulary growth during the study abroad experience was approximately four times faster than the rate observed in the learners' home country. However, it is important to note that these overall results concealed significant individual differences among the participants. While most students showed progress, one student's score remained unchanged, and five students obtained lower scores on the final test compared to their initial entry test (Kinginger,2009).

Sanz & Morales-Front (2018) investigated writing in SA, with special attention to those aspects of writing development most influenced by the immersive context. All the studies summarized here suggest that both explicit instruction in L2 writing and immersion abroad can help to improve L2 writing ability, but the specific aspects that are influenced by each vary.

Writing fluency

Just as speaking, writing proficiency can also be evaluated by fluency but with different measurement. Several interpretations of writing fluency exist, with qualitative definitions being common. For instance, Wolfe-Quintero et al. (1998) describe it as the ability to produce written language rapidly, appropriately, creatively, and coherently. Reynolds (2005) adds that writing fluency involves using linguistic structures effectively to fulfill rhetorical and social purposes. Alternatively, researchers adopting process-based definitions view writing fluency as encompassing the richness of writers' processes and their aptitude for organizing composing strategies (Bruton and Kirby, 1987). Additionally, Snellings et al. (2004) emphasize the speed of lexical retrieval during the act of writing as an important aspect of writing fluency(Abdel Latif, 2013, p. 101). Abdel Latif (2013, p. 101) 'pointed out that writing fluency can be measured by writers' pausing (Spelman Miller, 2000), changes made to the text (Knoch, 2007), composing rate (Sasaki, 2000), text quantity (Baba, 2009), length of translating episodes written between pauses (Abdel Latif, 2009), length of rehearsed text between pauses (Chenoweth & Hayes, 2001), linguistic features characterizing rhetorical functions (Reynolds, 2005), number and length of T-units (Storch, 2009), sentence length (Johnson et al., 2012), and text structure, coherence, and cohesion (Storch, 2009).

In the field of L2 writing fluency, several studies have examined the effects of study abroad experiences. Manchón& Polio (2021, p.262) pointed out that while some research findings suggest that there is no advantage for SA in terms of written fluency (Freed et al., 2004), others indicate that the benefits are not clear-cut (Llanez & Muñoz, 2013). Conversely, a number of studies have reported improvements in writing fluency following study abroad (Godfrey et al., 2004; Serrano et al., 2012; Sasaki, 2004, 2007; Pérez-Vidal & Barquin, 2014;).

Writing accuracy

Housen et al (2012) pointed out that accuracy, or correctness, essentially involves measuring how much an L2 learner's performance differs from a standard, typically represented by native speakers (Leow, 1993; Pallotti, 2009; Wolfe-Quintero et al., 1998). These variations from the standard are commonly referred to as 'errors' in Godfrey et al. (2004). For instance, it was found that there were more gendered language instances with fewer errors in gender agreement.

2.4 Diaries as data source

In L2 writing studies, written diaries were not as prevalent as other forms of writing, such as academic writing or elicited writing tasks, in serving as a data resource. The rationale for considering diaries suitable in this study for examining the formulaic sequences used in L2 writing output and exploring their development and influence lies in their ecological validity and authenticity. Moreover, diaries offer a valuable resource for tracing the origin of L2 learning, and they facilitate the easy recording and extraction of data.

Ecological validity refers to the extent to which the findings and claims of a study accurately represent real-world situations and settings. It involves assessing whether the conditions and variables studied in a research context closely mirror those encountered in the natural environment or everyday life. Achieving ecological validity is crucial for the generalizability and applicability of research findings to real-world scenarios (Cicourel, 2007). To generate the diaries that we apply in this study, no instruction was given about the requirement of formulaic sequences that the L2

Chinese learner should use in the diaries. They are free to apply any kinds of formulaic sequences and decided the quantity as they prefer. In this way, we can gain a high Ecological validity and authenticity.

Moreover, diaries can serve as a valuable resource for tracing the origin of L2 learning. While it may be challenging to pinpoint the exact time and source of new word acquisition during study abroad, the content of the diaries can provide valuable clues in understanding this process.

2.5 Summary

In this chapter, we thoroughly examined the definition and categorization of formulaic sequences. Chinese formulaic sequence (yǔkuài) is precisely defined as a 'prefabricated non-word sequence comprising two or more morphemes, limited to the length of a sentence, and often stored and retrieved as a cohesive unit during usage' (Wang, 2017). Through a comparative analysis of existing research, we systematically classified Chinese formulaic sequences into three types based on their form flexibility and functions. Subsequently, our exploration delved into the pivotal role that formulaic sequences play in the context of L2 Chinese learning during study abroad, elucidating their profound influence on writing skills. Lastly, we elucidated the rationale behind our selection of diaries as the primary data source for scrutinizing the aforementioned factors.

Chapter 3: Methodology

3.1 Introduction

In the preceding chapter, an overview of the definition and classification of formulaic sequences was provided, both in a general context and with a specific focus on formulaic sequences in Chinese. In order to observe changes in the frequency and correctness of formulaic sequences used across two periods, and furthermore, to examine the possible influence of formulaic sequence learning on L2 Chinese writing output development during study abroad experiences, in this chapter, I will explain how the formulaic sequences were extracted.The process of analysing formulaic sequences involved five main steps: preparation, extracting formulaic sequences, assessing their conventionality, evaluating the correctness of formulaic sequences extracted, and checking their correlation with the writing output which are measured by lexical diversity and mean sentence length.

3.2 Participants

The study aims to explore the use of formulaic sequences by L2 Chinese learners during their study abroad in the target country. To this end, 26 students who were studying Mandarin at a UK university were recruited as volunteers. These students were all in the second year of their studies, which required them to spend a year in a Chinese speaking country. The participants' language proficiency levels ranged from intermediate to upper-intermediate.

3.3 Data collection

Data for this study were collected through diaries written in Chinese, which participants were asked to create during their stay abroad as part of their final assessment. Two sets of diaries, each comprising two entries, were collected at different points during the participants' stay, and the time interval between the two sets of diaries is six months. The collected data were analysed to identify patterns and variations in the use of formulaic sequences by the participants, providing insights into the role of formulaic sequences in second language learning.

3.3 Research design of this study

The study utilised 104 diaries from the participants, with each participant contributing two diaries for each set. The diaries were all regular accounts of the students' daily routines, social activities, and interactions with family and friends. Participants were instructed to handwrite their diaries with no time limit, and each diary was expected to be around 700 words. The first set of diaries (diary 1 and 2) served as a baseline measure for writing proficiency, allowing us to observe any changes in writing proficiency after six months of further stay in target country.

In this study, we aim to answer the following three research questions:

1. What patterns can be seen in the usage of formulaic sequences according to their fixedness?

2. Was there any improvement of the usage of formulaic sequences, measured by frequency and accuracy, during the SA (from period 1 to period 2)?

3. Did the usage of the three types in of formulaic sequences, measured by frequency and accuracy, influence the lexical diversity and mean sentence length of the written texts respectively?

As mentioned in the Literature Review, the flexibility of formulaic sequence varies. To further understand their roles in SLA, we need to examine each type of formulaic sequence, from the most fixed form to the most free form. To extract three types of formulaic sequence from diaries (fixed, semi-fixed, free), the handwritten files were transcribed into a digital format and analysed using CLAN (MacWhinney, 2000) and Sketch Engine (Kilgarriff et al., 2004).

3.4 Procedure for extracting and analysing formulaic sequences

3.4.1 Preparation

Tools: CLAN and Sketch Engine

In this study, we use the CLAN program and Sketch Engine for extracting and analysing formulaic sequences, while SPSS[11] is employed for data analysis.

The CLAN (Computerized Language Analysis) program (MacWhinney, 2000) serves as a transcription editor, offering capabilities for transcription, coding, and analysis. It has linguistic analysis routines that can automatically analyse a basic transcript with precise morphological and syntactic annotations. CLAN supports multiple methods for transcribing audio and video data, facilitates searching, and enables data export and import. It is a valuable tool for deriving Language Sample Analysis (LSA) measures, such as MLU (Mean Length of Utterance)[12], lexical diversity (D-value and TTR (Token Type Ratio)), and various other LSA outcomes (MacWhinney, 2000)The data generated can be used for data analysis by SPSS. In our study, we work with written texts, so we skipped the audio transcription process. We only need to transform diaries into the CHAT format (MacWhinney, 2000), as required by CLAN, for analysing MLU and D-value.

To assess lexical diversity, we've opted for the D-value instead of TTR. This choice is driven by our use of a relatively small corpus consisting of only 104 diaries written by the participants. TTR calculates the ratio of unique words (types) to the total words (tokens) in a text, but this measure can be significantly influenced by the text's length. On the other hand, D-value, which is generated by VOCD within CLAN, takes into account the impact of sample size. This feature makes D-value a more reliable metric for comparing texts of different lengths (Duran et al., 2004). VOCD begins its analysis with 35 tokens and establishes the initial point on the transcript's curve by conducting 100 random trials, each involving the selection of 35 tokens from the text without replacement, followed by the computation of their average TTR. The choice of random sampling in VOCD is deliberate and serves to prevent potential reliability issues that may arise from clustering the same vocabulary items at specific points within the transcript (Duran et al., 2004).

Sketch Engine is a corpus tool which takes as input a corpus of any language(Kilgarriff & Kosem,2012) and generates word sketches of that language to illustrate words' grammatical and collocational behavior. It not only offers ready-to-

[11] SPSS (Statistical Package for the Social Sciences) is a software program widely used for statistical analysis in various fields, including social sciences, business, and health sciences. Developed by IBM, SPSS provides a comprehensive set of tools for data management and statistical (SPSS®13.0 Brief Guide, 2004).

[12] MLU (Mean Length of Utterance): An utterance in speaking is analogous to a sentence in writing; therefore, the mean sentence length can also be measured by MLU.

use corpora but also allows users to upload their own corpora in different languages for further analysis (Kilgarriff et al., 2014). To assess the correlation between two elements in a sequence, the MI score is commonly employed. It denotes the strength of a collocation and measures the level of non-randomness when two words co-occur (Hunston, 2002, p. 71). Sketch Engine provides the MI score for two words in the chosen collocation, enabling the analysis, including the extraction of concordance for target words with other parts of speech. The process involves calculating the MI score and generating word lists for each grammatical category, such as verbs and conjunctions, based on their frequency.

Sketch Engine allows users to upload their own coprus as well as offers the uploaded native speaker corpus in various language. For example, the Chinese corpus available on Sketch Engine is the *Chinese Web Corpus (zhTenTen)*,hereafter referred to as *zhTenTen*, which is composed of texts collected from the Internet. It is divided into simplified and traditional Chinese corpora. Although some of the participants wrote their diaries in traditional Chinese, we transformed them into simplified Chinese to ensure uniform data. We chose simplified Chinese for two primary reasons: first, because Chinese language courses at the UK university teach simplified Chinese, and second, because the difference between these two writing systems does not impact the extraction of the three types of formulaic sequences or the error analysis in this study.

The zhTenTen corpus on Sketch Engine is consisted of 13.5 billion words .The tools available to work with this corpus are:

word sketch– Chinese collocations categorized by grammatical relations
thesaurus– synonyms and similar words for every word
keywords– terminology extraction of one-word and multi-word units
word lists – lists of Chinese nouns, verbs, adjectives etc. organized by frequency
n-grams– frequency list of multi-word units
concordance – examples in context

Word lists and concordance functions with MI value indicated are used for extracting high-frequency V+N(object) high-frequency collocations from uploaded corpus (consisting of 52 corpora, each from the two sets of 26 participants). The concordance function and the collocation function are employed for checking conventionality and conducting error analysis in both the uploaded corpus and the *zhTenTen* corpus.

Data input in CLAN

To prepare the data for input into CLAN, the first step is to transcribe the diaries into CHAT format. This transcription process involves breaking down the texts into individual words to facilitate word counting and other data processing. Specifically, when dealing with the Chinese language, it's important to note that, unlike English, Chinese does not employ clear 'word boundaries' represented by spaces between words in written text. Therefore, written Chinese text needs to undergo a process known as 'Word Segmentation' (Xue & Shen, 2003) to mark those boundaries to obtain each individual word. This process is illustrated in the following example.

In English, words are written with clear boundaries: 'My_classmates_are_really_very_friendly.' (from diary set 1 of participant 2) However, in Chinese, the whole sentence is typically written without word segmentation: '我的同学真的很友好.' After applying word segmentation, the sentence will be marked as: '我_的_同学_真的_很_友好。'

The procedure of word segmentation, a critical step in this process, can be accomplished through the use of natural language processing (NLP) technology. In our study, we use the Language Technology Platform (LTP) (Che et al., 2010), accessible at http://ltp.ai/demo.html. LTP is a comprehensive Chinese language processing system developed by the Harbin Institute of Technology. LTP has pioneered the creation of an XML-based representation for language processing outcomes, which serves as the foundation for a suite of robust and high-performance Chinese language processing modules.

The next step involves editing the segmented diaries in CLAN and saving them as CHAT files. Subsequently, the morphology command in CLAN is applied to label each word with its corresponding morphological features. This allows CLAN to recognize, extract, and analyze words using various commands. Additionally, it can indicate the percentage of recognizable words, which proves valuable for validating the data in the study. After tagging, the texts will appear as demonstrated in the following example.

*PAR: 只要 我 将来 继续 上课 ， 久而久之 ， 听力 就 成为 潜意识

%mor: conj|zhi3yao4=so_long_as pro:per|wo3=I n:tm|jiang1lai2=future

v|ji4xu4=continue v|shang4ke4=go_to_classes cm|cm ?|久而久之 cm|cm
n|ting1li4=hearing adv|jiu4=just v|cheng2wei2=become
n|qian2yi4shi4=subconscious .

Morphology explanation**:**

The numbers in 'zhi3yao4' are used to indicate the tones in Chinese, so it is 'zhǐ yào.'.

1.'只要' (zhǐ yào)means 'so long as' and functions as a conjunction.

2. '我' (wǒ) means 'I' and functions as a personal pronoun.

3. '将来' (jiāng lái) means 'future' and functions as a noun of time.

4. '继续' (jì xù) means 'continue' and functions as a verb.

5. '上课' (shàng kè) means 'go to classes' and functions as a verb.

6. '久而久之' (jiǔ ér jiǔ zhī) , marked by a question mark '?' means that this word/phrase is not recognized in CLAN. It is actually an four-character idiom, meaning 'over time,' and is used to express the idea of something happening gradually or over an extended period.

7. '听力' (tīng lì) means 'hearing' and functions as a noun.

8. '就' (jiù) means 'just' and functions as an adverb.

9. '成为' (chéng wéi) means 'become' and functions as a verb.

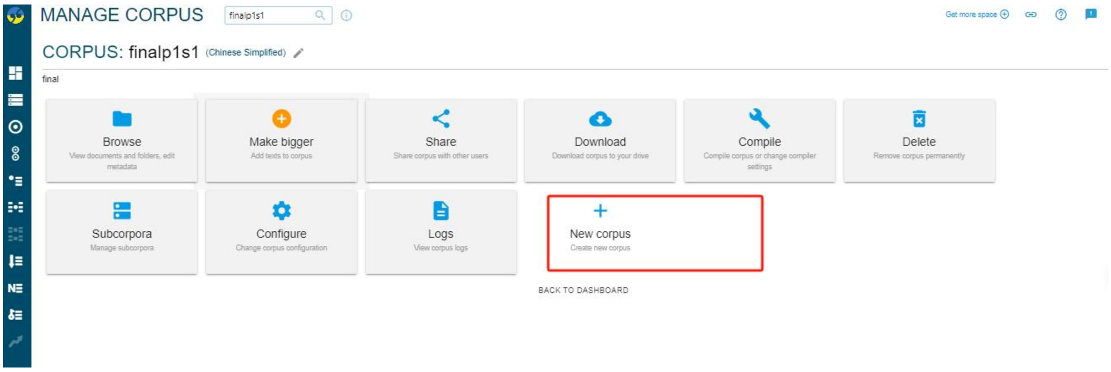
10. '潜意识' (qián yì shí) means 'subconscious' and functions as a noun.

The intervention of native speaker judgement (by the author) is another indispensable factor to ensure the utmost precision in word segmentation. This necessity arises from the potential for inaccuracies rooted in the limitations of the corpus employed for the training of language modules.

Uploading corpora to Sketch Engine

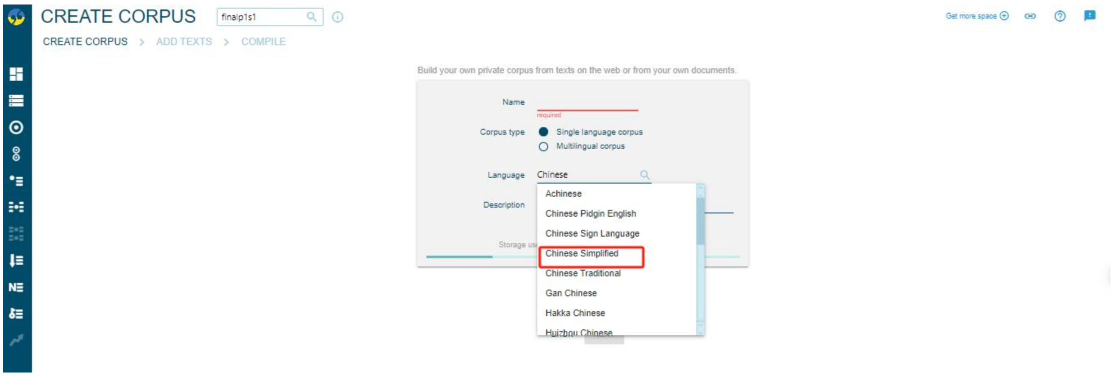
We can upload each set of diaries from every participant to Sketch Engine, naming them in the format 'finalp1s1' (indicating participant 1, set 1), for example. Each set will create a distinct corpus, as demonstrated in Figure 1. Subsequently, we can utilize any function available in SketchEngine (e.g. as displayed in the left column in figure 1) to conduct further analysis.

Figure 1 How to upload corpus onto Sketch Engine



CORPUS: finalp1s1 (Chinese Simplified)

final



Build your own private corpus from texts on the web or from your own documents.

Name
required

Corpus type ● Single language corpus ○ Multilingual corpus

Language Chinese

Description

3.4.2 Extracting formulaic sequences from CLAN and Sketch Engine

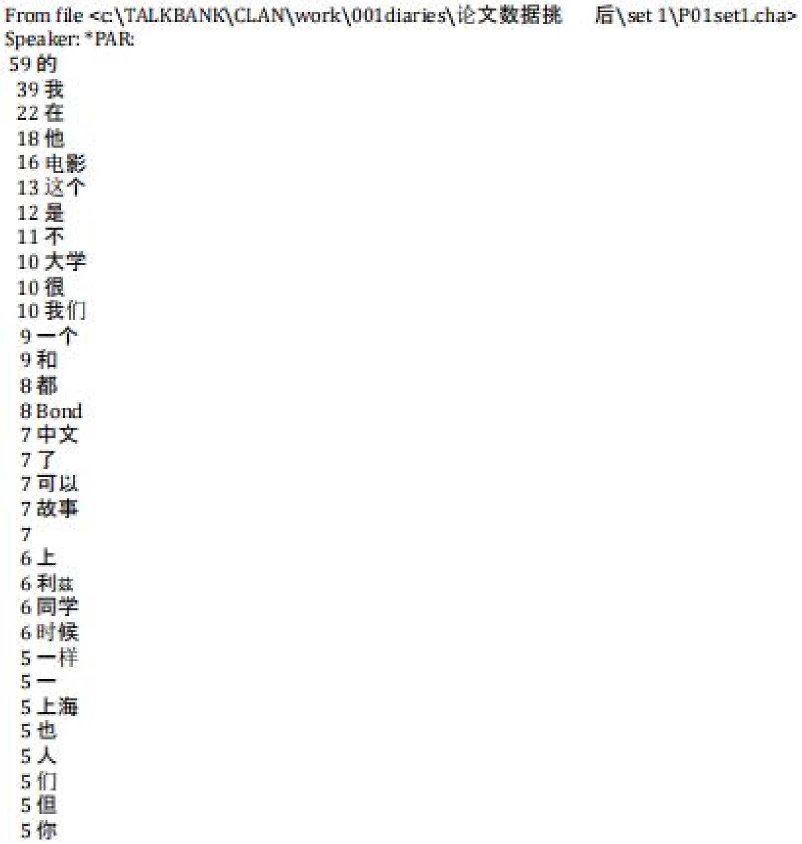
In this study, we extract the most fixed form of formulaic sequence, the four-character idiom, and the semi-fixed conjunctions from CLAN because it can apply the frequency descent command[13] for the former and precise morpheme tagging for the latter. For high-frequency V+N(object) high-frequency collocations, we turn to Sketch Engine as CLAN cannot highlight every sentence with verbs, which is necessary for our analysis. Furthermore, we can assess the MI value and validate these high-frequency V+N(object) high-frequency collocations using the native speaker corpora *zhTenTen*.

Extracting four-character idiom

Firstly, we need to extract all words composed of four characters. We employ the CLAN frequency descending command 'freq +t*PAR +o,' and we will obtain all the words with descending frequency in one set of the diary, such as the following Figure 2.

[13] Command 'freq +t*PAR +o' in ClAN can be used to extract all the words with a descending frequency in the selected file

Figure 2. Word List with Descending Frequency in Diary Set 1 by Participant 1 from CLAN.



```
From file <c:\TALKBANK\CLAN\work\001diaries\论文数据挑    后\set 1\P01set1.cha>
Speaker: *PAR:
 59 的
  39 我
  22 在
  18 他
  16 电影
  13 这个
  12 是
  11 不
  10 大学
  10 很
  10 我们
   9 一个
   9 和
   8 都
   8 Bond
   7 中文
   7 了
   7 可以
   7 故事
   7
   6 上
   6 利兹
   6 同学
   6 时候
   5 一样
   5 一
   5 上海
   5 也
   5 人
   5 们
   5 但
   5 你
```

We put the full list into Microsoft Excel to sort it by character count, resulting in a new word list ranked by the number of characters along with their frequency. For instance as Table 5 shows, in the first set of diaries written by Participant 1, the four-character words include *nírìlìyà* '尼日利亚'(Nigeria), *yīsībùgǔ* '一丝不苟' (meticulous, with every detail attended to), *jiǔ'érjiǔzhī* '久而久之' (over time, gradually)', *lìsuǒbùjí* '力所不及' (beyond one's capability, beyond one's reach), and mǎláixīyà '马来西亚 ' (Malaysia). However, *nírìlìyà* (Nigeria) and mǎláixīyà (Malaysia) are not idioms; they are simply country names. Consequently, we remove them from the list of four-character idioms.

Table 5. Word list of diaries written by participant 1 in Time 1, ranked by character count[14]

Word List	Frequency
RamiMalck	1
James	4
Bond-	1
Craig	1
craig	1
Bond	8
nírìlìyà '尼日利亚'(Nigeria)	2
yīsībùgǔ '一丝不苟' (meticulous, with every detail attended to)	1
jiǔ'érjiǔzhī '久而久之' (over time, gradually)'	1
lìsuǒbùjí '力所不及' (beyond one's capability, beyond one's reach)	1
mǎláixīyà '马来西亚 ' (Malaysia)	1
⋮	⋮

To further ascertain whether an idiom qualifies as a 'chengyu' (four-character idiom), we refer to the hànyǔ chéngyǔ zìdiǎn '汉语成语字典'(2004), which is a dictionary of four-character idioms that contains a comprehensive resource encompassing over 10,000 commonly used four-character idioms. This dictionary provides a wealth of information for each idiom, including its pronunciation in pīnyīn, definition, origin, historical anecdotes, example sentences, synonyms, near synonyms, antonyms, and distinctions. We will use this dictionary as a reference to check whether the four-character idioms that we extracted can be defined as 'chengyu'.

Extracting conjunctions

Since every word in CLAN has been morphologically tagged, to extract conjunctions, we just need to use the following command: *freq +t*PAR +d5 +o +sm;*,|conj*,o% +u *gem.cex.* Taking the two diaries from set 1 written by participant 1 as an example (Table 6.), we obtain the following list of conjunctions ranked by frequency. The total number of types and (tokens) is also available.

[14] The original data extracted from CLAN is in Chinese. Pingyin and English translations are added for better understanding. The Chinese data throughout this study will be represented mostly in pīnyīn (with Chinese characters and English explanation when necessary), for ease of readability.

Table 6. Conjunction list extracted from diaries of Participant 1 in Time 1[15]

```
1\P01set1.cha>
Speaker: *PAR:
  7 conj|he2 (和): and
  4 conj|suo3yi3 (所以): so
  3 conj|bi3ru2 (比如): for example
  3 conj|dan4 (但): but
  3 conj|sui1ran2 (虽然): although
  3 conj|yin1wei4 (因为): because
  2 conj|er2qie3 (而且): moreover
  1 conj|bu2dan4 (不但): not only
  1 conj|bu4guan3 (不管): no matter
  1 conj|chu2le (除了) : except
  1 conj|dan4shi4 (但是): however
  1 conj|huo4zhe3 (或者): or
  1 conj|ru2guo3 (如果): if
  1 conj|wei4le (为了): in order to
  1 conj|you2yu2 (由于): due to
  1 conj|zhi3yao4 (只要): only if
------------------------------
   16  Total number of different item types used
   34  Total number of items (tokens)
      0.471  Type/Token ratio
```

In this process, we need to check if the tags in CLAN are accurate, because certain Chinese characters, such as '和,' have the capacity to fulfil multiple functions, encompassing both conjunctions and adverbs. For this kind of words, it can be problematic for CLAN to mark the morphological features accuratly. For example, within the scope of our study, we only consider instances where *hé* '和'(and) functions as a conjunction, therefore we exclude the instances in which *hé* serves as a preposition.

[15] Conjunctions extracted from CLAN are indicated by pīnyīn. Chinese characters and English translations are added for etter understanding

Conversely, in the sentence '我同学们的年龄和生活情况都很有意思,' '和' indeed functions as a conjunction. Hence, the exercise of native speaker judgement is indispensable for the accurate identification and differentiation of instances in which ' 和' serves as a conjunction within the text.

Extracting V+N high frequency collocations

To extract high-frequency V+N(object) high-frequency collocations, on the contrary, it is more effective to use Sketch Engine than CLAN. If we compare Table 7 and Table 8, we can observe that from CLAN, only 94 types of verbs with 169 tokens are extracted. However, from the same diary, when using Sketch Engine, we obtain 118 types of verbs with 191 tokens, and all the verbs are valid by native speaker judgement.

Table 7. Verb list extracted from diaries written by participant 1 in Time 1 from CLAN

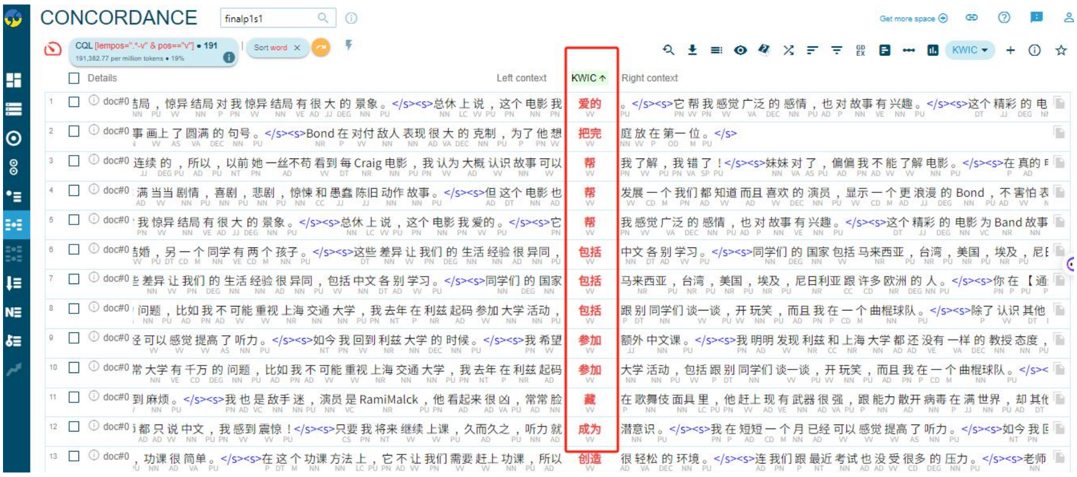
```
1\P01set1.cha>
Speaker: *PAR:
12 v:cop|shì '是' (be)
  7 v:aux|kěyǐ '可以' (can)
  7 v|yǒu '有' (have)
  5 v|zài '在' (at)
  4 v:dirc|qù '去' (go)
  4 v|liǎojiě '了解' (understand)
  4 v|shì '是' (is)
  4 v|shuō '说' (say)
  4 v|tīngdào '听到' (hear)
  3 v:dirc|shàng '上' (on)
  3 v|bāng '帮' (help)
  3 v|bāokuò '包括' (include)
  3 v|jìde '记得' (remember)
  3 v|jìxù '继续' (continue)
  3 v|juéde '觉得' (feel)
  3 v|kàn '看' (look)
  3 v|kàndào '看到' (see)
  3 v|ràng '让' (let)
        ⋮
------------------------------
   94  Total number of different item types used
  169  Total number of items (tokens)
0.556  Type/Token ratio
```

Table 8. Verb List extracted from diaries written by participant 1 in Time 1 from Sketch Engine

Verb List	Frequency
'shì' (是) (be)	11
'yǒu' (有) (have)	8
'kěyǐ' (可以) (can)	7
'yīyàng' (一样) (be the same)	5
'qù' (去) (go)	4
'tīngdào' (听到) (hear)	4
'shuō' (说) (say)	4
'méiyǒu' (没有) (do not have)	3
'jīngyì' (惊异) (be surprised)	3
'liǎojiě' (了解) (understand)	3
'ràng' (让) (let)	3
'bāokuò' (包括) (include)	3
'bāng' (帮) (help)	3
'yào' (要) (want)	3
'jìxù' (继续) (continue)	3
'juéde' (觉得) (feel)	3
'kàndào' (看到) (see)	3
'hǎo' (好) (good)	3
'jìde' (记得) (remember)	3
'xuéxí' (学习) (study, learn)⁝	2
	⋮
Total number of different item types used	118
Total number of items (tokens)	191

Another reason Sketch Engine is preferable is that it allows us to center and highlight verbs, along with other parts of speech, as shown in Figure 3.

Figure 3. Centred and Highlighted Verb List of Diary Set 1 by Participant 1 from Sketch Engine



CONCORDANCE finalp1s1

CQL [lempos=".*-v" & pos=="V"] • 191 | Sort word ×

191,382.77 per million tokens • 19%

Get more space

KWIC

	□ Details	Left context	KWIC ↑	Right context
1	□ doc#0	吉局，惊异 结局 对 我 惊异 结局 有 很 大 的 景象。</s><s>总休 上 说，这个 电影 我	爱的	。</s><s>它 帮 我 感觉 广泛 的 感情，也 对 故事 有 兴趣。</s><s>这个 精彩 的 电
2	□ doc#0	事 画上 了 圆满 的 句号。</s><s>Bond 在 对付 敌人 表现 很 大 的 克制，为了 他 想	把完	庭 放 在 第一 位。</s>
3	□ doc#0	连续 的，所以，以前 她 一丝不苟 看到 每 Craig 电影，我 认为 大概 认识 故事 可以	帮	我 了解，我 错 了！</s><s>妹妹 对 了，偏偏 我 不 能 了解 电影。</s><s>在 真的 F
4	□ doc#0	满 当当 剧情，喜剧，悲剧，惊悚 和 愚蠢 陈旧 动作 故事。</s><s>但 这个 电影 也	帮	发展 一 个 我们 都 知道 而且 喜欢 的 演员，显示 一 个 更 浪漫 的 Bond，不 害怕 表
5	□ doc#0	我 惊异 结局 有 很 大 的 景象。</s><s>总休 上 说，这个 电影 我 爱的。</s><s>它	帮	我 感觉 广泛 的 感情，也 对 故事 有 兴趣。</s><s>这个 精彩 的 电影 为 Band 故事
6	□ doc#0	吉婚，另 一 个 同学 有 两 个 孩子。</s><s>这些 差异 让 我们 的 生活 经验 很 异同，	包括	中文 各 别 学习。</s><s>同学们 的 国家 包括 马来西亚，台湾，美国，埃及，尼E
7	□ doc#0	差异 让 我们 的 生活 经验 很 异同，包括 中文 各 别 学习。</s><s>同学们 的 国家	包括	马来西亚，台湾，美国，埃及，尼日利亚 跟 许多 欧洲 的 人。</s><s>你 在【通
8	□ doc#0	问题，比如 我 不 可能 重视 上海 交通 大学，我 去年 在 利兹 起码 参加 大学 活动，	包括	跟 别 同学们 谈一谈，开 玩笑，而且 我 在 一 个 曲棍球队。</s><s>除了 认识 其他
9	□ doc#0	经 可以 感觉 提高 了 听力。</s><s>如今 我 回到 利兹 大学 的 时候。</s><s>我 希望	参加	额外 中文课。</s><s>我 明明 发现 利兹 和 上海 大学 都 还 没有 一样 的 教授 态度，
10	□ doc#0	常 大学 有 千万 的 问题，比如 我 不 可能 重视 上海 交通 大学，我 去年 在 利兹 起码	参加	大学 活动，包括 跟 别 同学们 谈一谈，开 玩笑，而且 我 在 一 个 曲棍球队。</s><
11	□ doc#0	到 麻烦。</s><s>我 也 是 敌手 迷，演员 是 RamiMalck，他 看起来 很 凶，常常 脸	藏	在 歌舞伎 面具 里，他 赶上 现 有 武器 很 强，跟 能力 散开 病毒 在 满 世界，却 其他
12	□ doc#0	都 只 说 中文，我 感到 震惊！</s><s>只要 我 将来 继续 上课，久而久之，听力 就	成为	潜意识。</s><s>我 在 短短 一 个 月 已经 可以 感觉 提高 了 听力。</s><s>如今 我 回
13	□ doc#0	，功课 很 简单。</s><s>在 这 个 功课 方法 上，它 不 让 我们 需要 赶上 功课，所以	创造	很 轻松 的 环境。</s><s>连 我们 跟 最近 考试 也 没 受 很多 的 压力。</s><s>老师

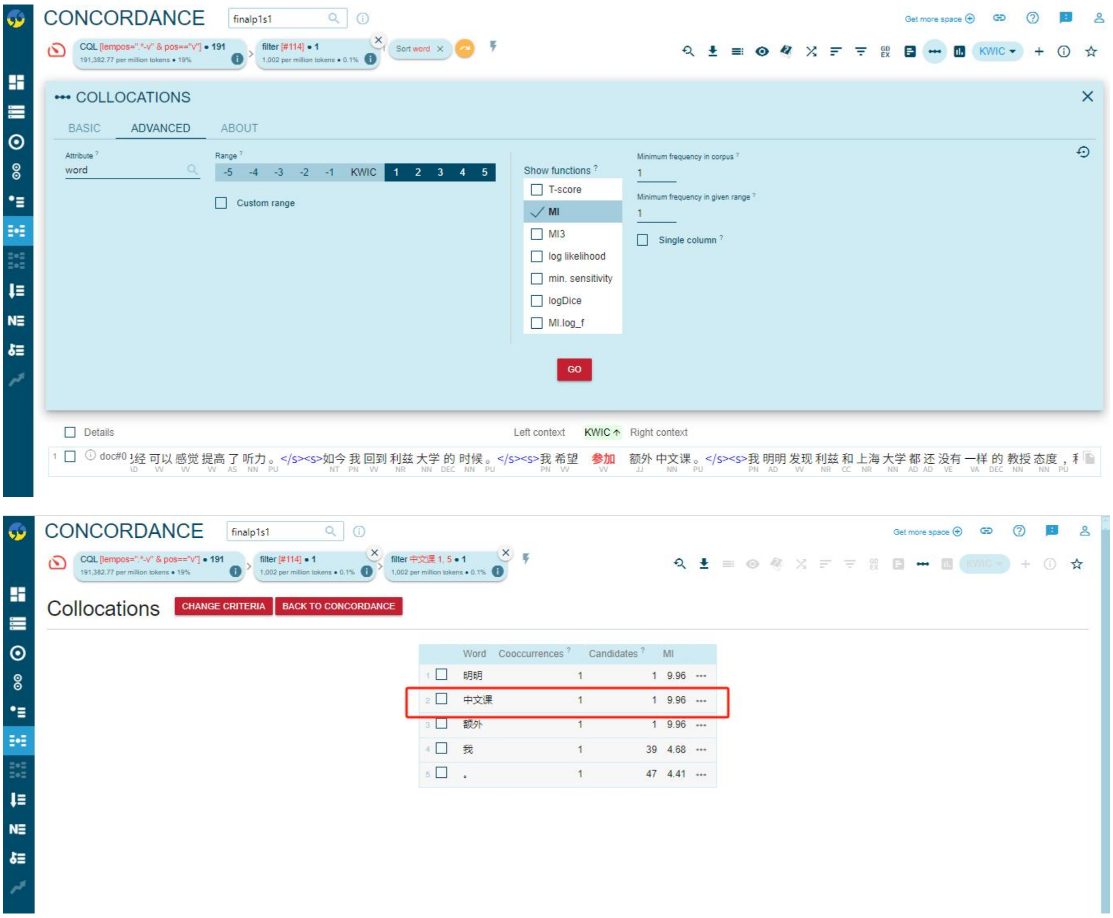
From the Sketch Engine corpora, we compile the list of verbs. In the subsequent step, our objective is to identify verbs that are followed by a noun or a noun phrase, where the noun functions as an object. This doesn't necessarily require that the noun be adjacent to the verb. Pronouns are included in this selection because they serve as nouns when functioning as objects. Before applying these criteria, we exclude copula verbs since they do not establish the conventional predicate-object relationship. Modal verbs are also omitted from consideration, as they are typically used in conjunction with other verbs rather than as objects. Additionally, verbs such as *juédé* '觉得' (think) and *rènwéi* '认为' (believe) are excluded because they are usually followed by complete sentences or clauses, rather than simple noun phrases.For example, in the sentence *zǒngtǐ shàng shuō, zhège diànyǐng wǒ ài de* '总体上说,这个电影我爱的' (generally speaking, I love this film) the verb '爱' (love) is followed by '的' (of), which is not a noun (object). Therefore, this expression is not categorized as a V+N(object) high-frequency collocations. Similarly, in the sentence *dàn zhège diànyǐng yě* ***bāng fāzhǎn*** *yī gè wǒmen dōu zhīdào érqiě xǐhuān de yǎnyuán*'但这个电影也**帮发展**一个我们都知道而且喜欢的演员' (but this movie also helps develop an actor that we all know and like), it does not qualify as a valid V+N(object) high-frequency collocations because it comprises a V+V structure, specifically '*bāng*' (help) and '*fāzhǎn*' (develop). Moreover, the verb *bǎ* '把 ' (to make; to let) and its collocations are not selected since on Sketch Engine, it is classified as one separate category instead one verb.

By following these criteria, we compile a comprehensive list of verbs for each participant, which will subsequently undergo further evaluation in subsequent stages.

Assessing the validation of V+N(object) collocation extracted by Sketch Engine

To determine whether a combination of words can be considered a collocation, both Mutual Information (MI) score and T-scores are widely used. In this study, the type of formulaic sequence that we need to check the strength of co-occurrence is high-frequency V+N(object) high-frequency collocations because the other two types of formulaic sequences are measured differently. We choose the MI score over the T-score for assessing V+N(object) high-frequency collocations primarily because of the size of our corpus. The MI score 'indicates the strength of a collocation' and 'measures the amount of non-randomness present when two words co-occur' (Hunston, 2002, p. 71). It can be effectively applied to small corpora. For example, even for rare words with low frequencies, such as 'baleful' and 'gaze,' the collocation between these two words can yield a high MI score. Therefore, the MI score is suitable for our diaries corpus uploaded to Sketch Engine. Process can be seen in Figure 4.

Figure 4. MI score of "参加" (attend) + "中文课"(Chinese language course) by concordance on Sketch Engine



CONCORDANCE finalp1s1

CQL [lempos=".*-v" & pos=="V"] • 191
191,382.77 per million tokens • 19%

filter [#114] • 1
1,002 per million tokens • 0.1%

Sort word

COLLOCATIONS

BASIC ADVANCED ABOUT

Attribute: word

Range: -5 -4 -3 -2 -1 KWIC 1 2 3 4 5

□ Custom range

Show functions
□ T-score
✓ MI
□ MI3
□ log likelihood
□ min. sensitivity
□ logDice
□ MI.log_f

Minimum frequency in corpus: 1

Minimum frequency in given range: 1

□ Single column

GO

□ Details

Left context KWIC Right context

1 □ doc#0]经 可以 感觉 提高 了 听力 。</s><s>如今 我 回到 利兹 大学 的 时候 。</s><s>我 希望 **参加** 额外 中文课 。</s><s>我 明明 发现 利兹 和 上海 大学 都 还 没有 一样 的 教授 态度 ，

CONCORDANCE finalp1s1

CQL [lempos=".*-v" & pos=="V"] • 191
191,382.77 per million tokens • 19%

filter [#114] • 1
1,002 per million tokens • 0.1%

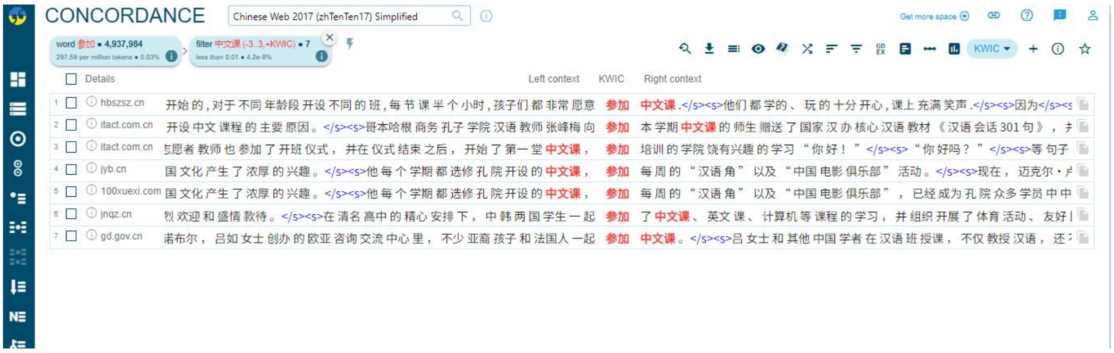
filter 中文课 1, 5 • 1
1,002 per million tokens • 0.1%

Collocations CHANGE CRITERIA BACK TO CONCORDANCE

		Word	Cooccurrences	Candidates	MI
1	□	明明	1	1	9.96
2	□	中文课	1	1	9.96
3	□	额外	1	1	9.96
4	□	我	1	39	4.68
5	□	。	1	47	4.41

Nevertheless, relying solely on the MI score does not always provide 'a reliable indication of a meaningful association' (Hunston, 2002, p. 72). For instance, the odd combination 'suprising hardly' may occasionally yield a higher MI score than the correct combination 'hardly surprising' (7.8) (Hunston, 2002). As a result, additional measures are required to further define high-frequency V+N(object) high-frequency collocations in our specific corpus by L2 Chinese learners. We employed zhTenTen to check whether a given collocation can be found within this extensive native speaker corpus. The search for the co-occurrence of these combinations can be done by using the filtering function of Sketch Engine (Kilgarriff et al., 2014) as indicated in Figure 5.

Figure 5. Concordance of "参加" (attend) + "中文课"(Chinese language course) collocation on Sketch Engine from Corpus zhTenTen 2017



Chinese Web 2017 (zhTenTen17) Simplified

word 参加 • 4,937,984
297.59 per million tokens • 0.03%

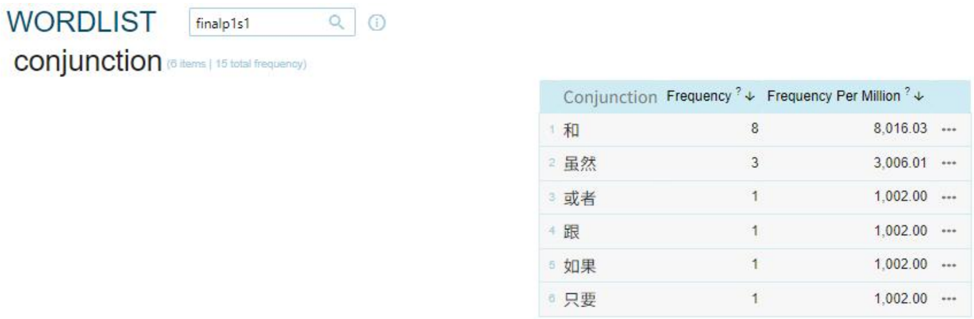
filter 中文课 (-3..3,+KWIC) • 7
less than 0.01 • 4.2e-8%

	□ Details	Left context	KWIC	Right context
1	□ hbszsz.cn	开始 的 , 对于 不同 年龄段 开设 不同 的 班 , 每 节 课 半 个 小时 , 孩子们 都 非常 愿意	参加	中文课 .</s><s>他们 都 学的 、 玩 的 十分 开心 , 课上 充满 笑声 .</s><s>因为</s><s
2	□ itact.com.cn	开设 中文 课程 的 主要 原因 。</s><s>哥本哈根 商务 孔子 学院 汉语 教师 张峰梅 向	参加	本 学期 中文课 的 师生 赠送 了 国家 汉 办 核心 汉语 教材 《 汉语 会话 301 句 》 ，
3	□ itact.com.cn	愿者 教师 也 参加 了 开班 仪式 ， 并在 仪式 结束 之后 ， 开始 了 第一 堂 中文课 ，	参加	培训 的 学院 饶有兴趣 的 学习 “ 你 好 ！ ” </s><s> “ 你 好吗 ？ ” </s><s>等 句子
4	□ jyb.cn	国 文化 产生 了 浓厚 的 兴趣 。</s><s>他 每 个 学期 都 选修 孔 院 开设 的 中文课 ，	参加	每 周 的 “ 汉语 角 ” 以及 “ 中国 电影 俱乐部 ” 活动 。</s><s>现在 ， 迈克尔 •
5	□ 100xuexi.com	国 文化 产生 了 浓厚 的 兴趣 。</s><s>他 每 个 学期 都 选修 孔 院 开设 的 中文课 ，	参加	每 周 的 “ 汉语 角 ” 以及 “ 中国 电影 俱乐部 ” ， 已经 成为 孔 院 众多 学员 中 中
6	□ jnqz.cn	烈 欢迎 和 盛情 款待 。</s><s>在 清名 高中 的 精心 安排 下 ， 中 韩 两 国 学生 一起	参加	了 中文课 、 英文 课 、 计算机 等 课程 的 学习 ， 并 组织 开展 了 体育 活动 、 友好
7	□ gd.gov.cn	诺布尔 ， 吕如 女士 创办 的 欧亚 咨询 交流 中心 里 ， 不少 亚裔 孩子 和 法国人 一起	参加	中文课 。</s><s>吕 女士 和 其他 中国 学者 在 汉语 班 授课 ， 不仅 教授 汉语 ， 还

3.4.3 Examination of frequency of four-character idioms, conjunctions and V+N(object) high frequency collocation.

To determine the frequency of different types of formulaic sequences, we utilize various tools. CLAN is chosen for four-character idioms and conjunctions because it allows for easy selection of four-character idioms by ranking word length, whereas extracting them from Sketch Engine is challenging. Regarding conjunctions, both CLAN and Sketch Engine can extract this type of formulaic sequence. However, Sketch Engine recognizes significantly fewer conjunctions compared to CLAN. As illustrated in Figure 6, in the same selected diary, Sketch Engine lists only 6 types of conjunctions, totaling 15 tokens. In contrast, CLAN extracts 16 types with 34 tokens, all of which are valid conjunctions, except for 'he,' which requires verification, as mentioned above.

Figure 6. Word List of Conjunctions Diary Set 1 by Participant 1 from Sketch Engine



WORDLIST finalp1s1

conjunction (6 items | 15 total frequency)

	Conjunction	Frequency ? ↓	Frequency Per Million ? ↓	
1	和	8	8,016.03	•••
2	虽然	3	3,006.01	•••
3	或者	1	1,002.00	•••
4	跟	1	1,002.00	•••
5	如果	1	1,002.00	•••
6	只要	1	1,002.00	•••

To determine the frequency of four-character idioms and conjunctions, we employ CLAN and utilize its 'freq@' command. This software allows us to extract frequency information both by token and by type. The token-based frequency measurement indicates the quantity of formulaic sequences in each selected file or folder, while the type-based measurement provides insights into the diversity. This diversity measurement is valuable for assessing the differential usage of formulaic sequences across the two periods.

By utilizing the CLAN extracting command 'req +tPAR +d5 +o +sm;,|conj*,o% +u *gem.cex,' we can generate lists of conjunctions used by each participant, considering both token and type frequencies. Consolidating these individual lists enables us to compile the overall frequencies of conjunctions.

In contrast to the extraction of four-character idioms and conjunctions by CLAN, we obtained high-frequency V+N(object) collocations from Sketch Engine using the uploaded corpus of diaries written by participants. The process began by listing verbs and then evaluating each one to determine if it could form a collocation with a noun (object). To qualify as a collocation, a specific frequency threshold needed to be met. As detailed in the Methodology section, we specifically chose collocations with a MI score of ≥3. The MI scores of each verb with other elements in each sentence are all provided within Sketch Engine. Subsequently, we compared the results with the native Chinese speaker corpus ZhTenTen from Sketch Engine to ascertain whether the V+N(object) high-frequency collocations could also be identified in native speaker speech. This step aims to address issues arising from the frequent use of unique collocations by L2 speakers. These procedures culminate in the compilation of the frequencies of V+N(object) high-frequency collocations used in Time 1 and 2.

After the selection and validation check of V+N(object) high-frequency collocations, we use conduct descriptive statistic data analysis to compare the difference across two periods.

3.4.4 Error analysis

Error analysis will be conducted by a qualitative approach, assessing whether the participants have used the selected types of formulaic sequences correctly and to identify any signs of improvement. These errors encompass both syntactic and lexical aspects. The evaluation primarily relies on native speaker judgement, conducted by the author, and searches within the Sketch Engine Chinese corpus to clarify uncertain errors. It's important to note that the absence of formulaic sequences is not categorized as an error, as this study specifically concentrates on the formulaic sequences actively used by the participants.

A quantitative approach will also be applied through descriptive statistics analysis in SPSS to compare the correctness of two periods. Additionally, we will explore whether there are observable improvements through this accuracy analysis in the use of each type of formulaic sequence over time during the study abroad period.

3.4.5 Examining the correlation of the use of formulaic sequences and writing output by SPSS[16]

We conducted a quantitative analysis of the frequency of each type of frequent sequences (formulaic sequences) used in Period 1 and Period 2 using SPSS[17] . Descriptive statistics is applied to calculate mean values and standard deviations to assess any significant differences of frequency in Time 1 and Time 2 by SPSS.

The evaluation of formulaic sequence accuracy across the two periods involved both quantitative and qualitative approaches. Initially, the accuracy of each formulaic sequence was individually assessed, and the data were then collectively examined. Subsequently, we computed the mean accuracy of each set of diaries for both periods. To explore the relationship between the number of each type of formulaic sequence

[17] SPSS (Statistical Package for the Social Sciences) is a software program widely used for statistical analysis in various fields, including social sciences, business, and health sciences. Developed by IBM, SPSS provides a comprehensive set of tools for data management and statistical (SPSS®13.0 Brief Guide, 2004).

used in the two periods and the changes in lexical diversity and mean sentence length produced by the participants, we conducted an linear regression analysis by SPSS.

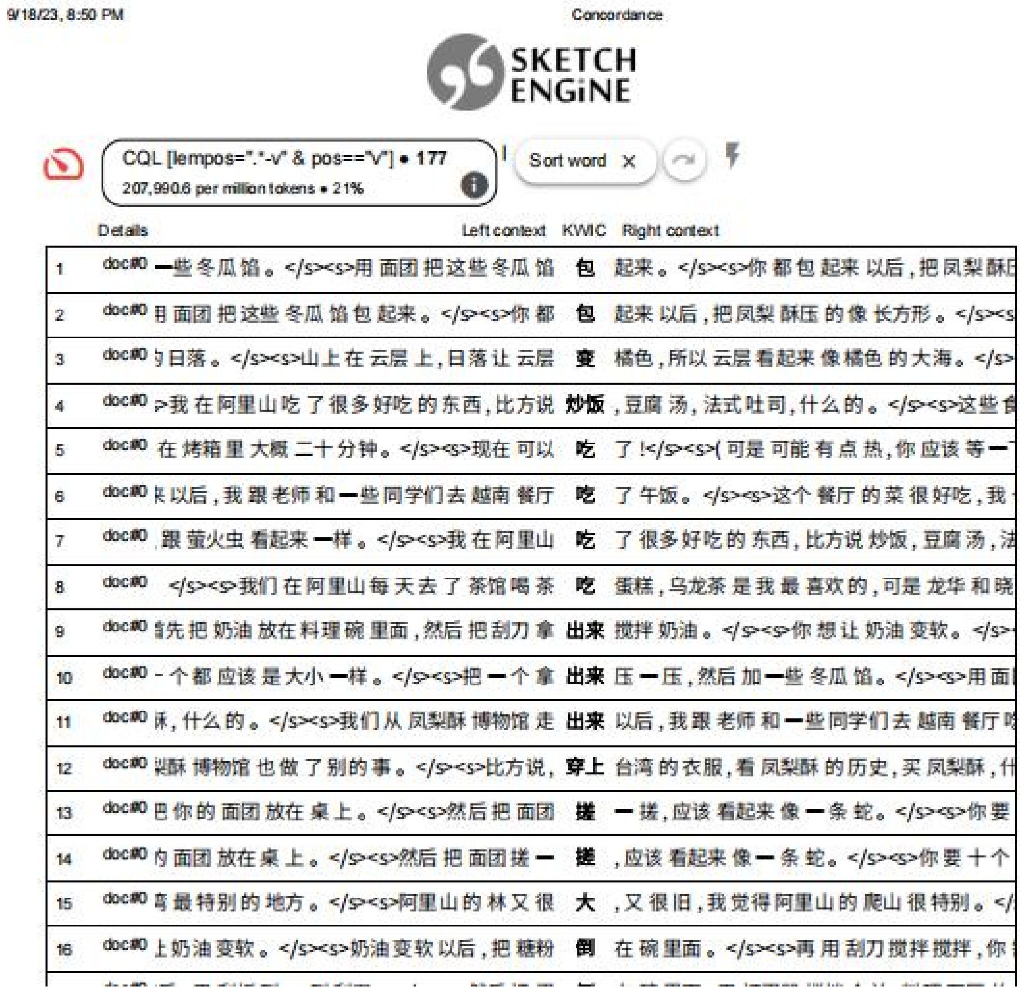
3.4.6 Examining reliability

Recognizing the significance of reliability and validity in qualitative research (Golafshani, 2015), our study, delving into the accuracy of formulaic sequences across two periods, adopted a qualitative approach. Given the inherent nature of subjective interpretation and a contextual focus in this type of research, the traditional notions of reliability and validity necessitate a redefinition. To ensure the trustworthiness of our findings, we enlisted native English speaker co-raters for judgements on acceptability and reliability. This methodological choice aligns with the nuanced demands of qualitative inquiry, emphasizing the adaptability of traditional research concepts to maintain the rigor and credibility of the study.

Among the three types of formulaic sequences, extracting V+N(object) high-frequency collocations poses the most challenges. One issue arises when a verb is not immediately followed by a noun, making it difficult for Sketch Engine to extract them as a collocation. Additionally, even when a verb is directly followed by a noun, it's crucial to verify if the noun functions as the object of the target verb, as L2 learners might construct sentences with incorrect word order. To ensure reliability, two Chinese native co-raters, both with backgrounds in linguistics—one as a university Chinese language course instructor and the other a Ph.D. candidate specializing in linguistics—were enlisted.

The co-raters were relieved from the need to consider the MI score of frequency since it was readily available on Sketch Engine. Their task solely involved extracting these collocations from randomly selected diaries. The instructor assessed 10% of the total data, while the Ph.D. candidate evaluated 5%. Both received detailed instructions and samples before commencing their assessments. In the randomly selected diaries, the verb list was extracted from Sketch Engine, with the verbs highlighted in bold, as illustrated in Figure 7.

Figure 7. Verb list with concordance from Sketch Engine.



9/18/23, 8:50 PM

Concordance

CQL [lempos=".*-v" & pos=="V"] • 177

207,990.6 per million tokens • 21%

	Details	Left context	KWIC	Right context
1	doc#0	一些 冬瓜 馅。</s><s>用 面团 把 这些 冬瓜 馅	包	起来。</s><s>你 都 包 起来 以后，把 凤梨 酥
2	doc#0	面团 把 这些 冬瓜 馅 包 起来。</s><s>你 都	包	起来 以后，把 凤梨 酥压 的 像 长方形。</s><s
3	doc#0	日落。</s><s>山上 在 云层 上，日落 让 云层	变	橘色，所以 云层 看起来 像 橘色 的 大海。</s>
4	doc#0	>我 在 阿里山 吃 了 很多 好吃 的 东西，比方说	炒饭	，豆腐 汤，法式 吐司，什么 的。</s><s>这些
5	doc#0	在 烤箱 里 大概 二十 分钟。</s><s>现在 可以	吃	了 !</s><s>(可是 可能 有 点 热，你 应该 等 一
6	doc#0	以后，我 跟 老师 和 一些 同学们 去 越南 餐厅	吃	了 午饭。</s><s>这个 餐厅 的 菜 很 好吃，我
7	doc#0	跟 萤火虫 看起来 一样。</s><s>我 在 阿里山	吃	了 很多 好吃 的 东西，比方说 炒饭，豆腐 汤，法
8	doc#0	</s><s>我们 在 阿里山 每 天 去 了 茶馆 喝 茶	吃	蛋糕，乌龙茶 是 我 最 喜欢 的，可是 龙华 和 晓
9	doc#0	先 把 奶油 放在 料理 碗 里面，然后 把 刮刀 拿	出来	搅拌 奶油。</s><s>你 想 让 奶油 变软。</s>
10	doc#0	- 个 都 应该 是 大小 一样。</s><s>把 一 个 拿	出来	压 一 压，然后 加 一些 冬瓜 馅。</s><s>用 面
11	doc#0	，什么 的。</s><s>我们 从 凤梨酥 博物馆 走	出来	以后，我 跟 老师 和 一些 同学们 去 越南 餐厅
12	doc#0	酥 博物馆 也 做 了 别的 事。</s><s>比方说，	穿上	台湾 的 衣服，看 凤梨酥 的 历史，买 凤梨酥，什
13	doc#0	你 的 面团 放在 桌 上。</s><s>然后 把 面团	搓	一 搓，应该 看起来 像 一 条 蛇。</s><s>你 要
14	doc#0	面团 放在 桌 上。</s><s>然后 把 面团 搓 一	搓	，应该 看起来 像 一 条 蛇。</s><s>你 要 十 个
15	doc#0	最 特别 的 地方。</s><s>阿里山 的 林 又 很	大	，又 很 旧，我 觉得 阿里山 的 爬山 很 特别。</
16	doc#0	上 奶油 变软。</s><s>奶油 变软 以后，把 糖粉	倒	在 碗 里面。</s><s>再 用 刮刀 搅拌 搅拌，你

The results show that the reliability rate is around 90% on average. The 10% discrepancy is mainly because verbs used as attributive for nouns, such as 'zuò de fàn' (cook^{V}-of-meal^{N}), were not included in our study since we chose only nouns as objects, but they were selected by the co-raters.

3.4.6 Ethical considerations

Data collection in this study commenced only after obtaining ethical approval from the University of Leeds Ethics Committee. The diaries were written two years ago and therefore do not involve the collection of live or real-time data. The participants' privacy and confidentiality have been rigorously maintained throughout the research process, and all data has been anonymised to ensure their identities are protected. In the analysis and reporting of results, no personally identifiable information will be disclosed, and any direct or indirect references to specific individuals will be avoided to uphold the highest ethical standards in research

3.5 Summary

In this chapter, we explained how to extract three types of formulaic sequences and get prepared for the data analysis and how to carry out the data analysis, as well as checking the validation of data. The methods are summarised in the following Table 9.

Table 9. Summary of the methods and tools for the extraction, anlysis and vadiliation of three types of formulaic sequences in this study.

Methods & tools		Fixed-form: four-character idiom	Semi-fixed form: conjunction	Free form: V+N(object) high frequency collocation
Tools to extract data		CLAN	CLAN	Sketch Engine
Methods to analysis data	Frequency	Quantitative approach aiming to compare individual development and overall group change in Time 1 and Time 2		
	Accuracy	Both qualitative and quantitative approaches are applied, with a focus on the qualitative aspect, which is carried out by native speaker judgement by the author.		
	Correlation to MLU and lexical diversity	Qualitative approach: examining the correlation by Wilcoxon	Qualitative approach: examining the correlation by paired sample T-test	
Tools to check the validation		Chinese four-character idiom dictionary	CLAN morphology tag& native speaker correction by the author	Two native Chinese co-raters

Our aim is to determine whether there is an increase in the frequency of using formulaic sequences during the sojourn in the target country. Additionally, we seek to assess the accuracy of formulaic sequences of these three types in two periods to explore what the participants have learned and identify potential challenges hindering their development in formulaic sequences learning. Furthermore, we aim to examine how the usage of formulaic sequences impacts writing output in terms of lexical diversity and mean sentence length among L2 Chinese learners.

Chapter 4 : Data Analysis

4.1 Introduction

In the previous chapter, I outlined the process of collecting and extracting all three types of formulaic sequences. In this chapter, I will investigate the frequency and accuracy of each type of formulaic sequence and explore the potential role these sequences can play in writing outputs, aiming a to address the research questions presented in Chapter 3.

Section 4.2 offers a comprehensive explanation of the methodology employed in conducting the frequency analysis for each type of formulaic sequence within the diaries. This process includes examining the frequency of each type used by each participant to discern patterns applied in their diaries, providing insights to address RQ 1. Furthermore, a comparison will be made between the overall, group, and individual frequencies of all three types of formulaic sequences at both Time 1 and Time 2. This comparative analysis aims to determine if any improvements were gained, addressing the frequency aspect of RQ2.

Section 4.3 investigates whether there are any changes in accuracy across the two periods, as posed by RQ2, by assessing the correctness of formulaic sequences used by participants in both Time 1 and Time 2 and analysing the errors. Through this analysis of errors, we can comprehend the challenges associated with each type of formulaic sequence and identify areas of improvement. This examination enables us to discuss the reasons behind these errors and, consequently, offer valuable suggestions for further studies in second language learning.

Section 4.4 explores whether the frequency of formulaic sequences used in each diary exerts any influence on the lexical and syntactical complexity of writing outputs. This is measured by analysing lexical diversity and mean sentence length, addressing Research Question 3.

4.2 Analysing the frequency of formulaic sequences used in Time 1 and Time 2

4.2.1 Frequency of four-character idioms

Following the procedures outlined in the methodology section, we ran the 'freq@' command in CLAN to extract all the words with frequencies in all the diaries in Time

1 and in Time 2 separately. We obtained the overall frequency list by token and ranked the words by the number of occurrences. Subsequently, we selected the four-character idioms which can be found in the 'Dictionary of Four-Character Idioms' (2004) and compiled the full list of four-character idioms by token, as shown in Appendix 1.

Based on the data presented in Appendix 1, it is evident that the idiom 'dà kāi yǎn jiè' was used twice in Time 1, while the remaining idioms were each used only once during that period. In Time 2, 'dà kāi yǎn jiè' (大开眼界) continued to be the most frequently used idiom which was applied four times. Several other idiom phrases, including 'rén shān rén hǎ'(人山人海), 'dú yī wú èr' (独一无二), 'liú lián wàng fǎn' (流连忘返), 'tí xīn diào dǎn' (提心吊胆) and 'zǒng ér yán zhī' (总而言之) were also employed multiple times. In addition to 'dà kāi yǎn jiè' (打开眼界) there were other idiom phrases that were used in both periods, such as 'liú lián wàng fǎn' (流连忘返) and 'rén shān rén hǎi' (人山人海)but with an increased frequency in Time 2, which shows a preference for specific idioms among the participants and a strengthening of this preference in Time 2.

To explore the frequency of idioms used by each participant, we ran the 'freq@' command in CLAN for each set of diaries of every participant. Subsequently, we collected the data, which allowed us to compile Chart 1, illustrating the quantity of idioms used in both Time 1 and Time 2 by token. Furthermore, we employed this data to conduct a statistical comparison, shown in Table 3.

Chart 1. Frequency of four-character idioms used by each participant in Time 1 and Time 2, measured by token.

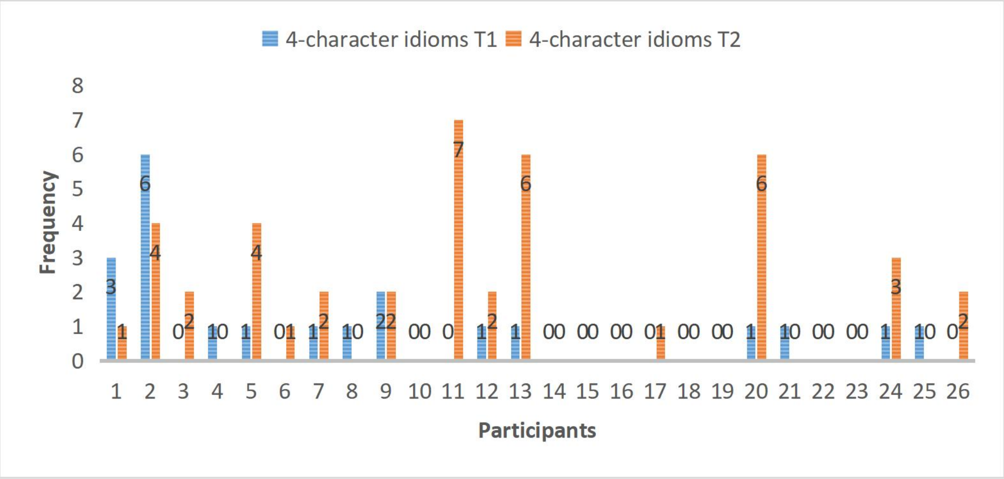


Chart 1 illustrates a noteworthy surge in the utilization of four-character idioms among participants. In Time 1, only 13 participants used this type of idiom, while in Time 2, the number increased to 23. Moreover, during Time 1, merely two students employed more than two idioms in each set of their diary; however, in Time 2, this figure expanded to 6. It means that more participants started to employ four-character idioms and with higher quantity in Time 2.

Table 10. Summary of the change of frequency of four-character idioms across the two periods.

	Minimum (number of participants)	Maximum (number of participants)	Sum	Mean	Std. Deviation
4-character idiomsT1	0 (13)	6.00 (1)	21	0.8077	1.29674
4-character idiomsT2	0 (3)	7.00 (1)	43	1.6538	2.13433

Table 10 statistically demonstrates a change in the usage of four-character idioms, revealing a substantial increase in the overall frequency of idiom usage. Instances rose from 21 in Time 1 to 43 in Time 2, nearly doubling in frequency. This significant increase is also highlighted by the doubling of the mean value. Additionally, the standard deviation exhibited a slight shift, moving from 1.29 in Time 1 to 2.13 in Time 2. This implies that in Time 2, participants not only tended to employ more idiom but showed a higher level of consistency in their idiom utilization.

4.2.2 Frequency of conjunctions

Utilizing the CLAN extracting command 'req +tPAR +d5 +o +sm;,|conj*,o% +u *gem.cex,' enables us to generate a list of conjunctions used by each participant, both by token and by type. Upon consolidating these individual lists, we compiled the frequencies of conjunctions, presented in both Chart 2 (by token) and Chart 3 (by type), and ranked them based on the increase in frequency from Time 1 to Time 2, moving from the most to the least increased.

Chart 2. Difference in the frequency of conjunctions by token

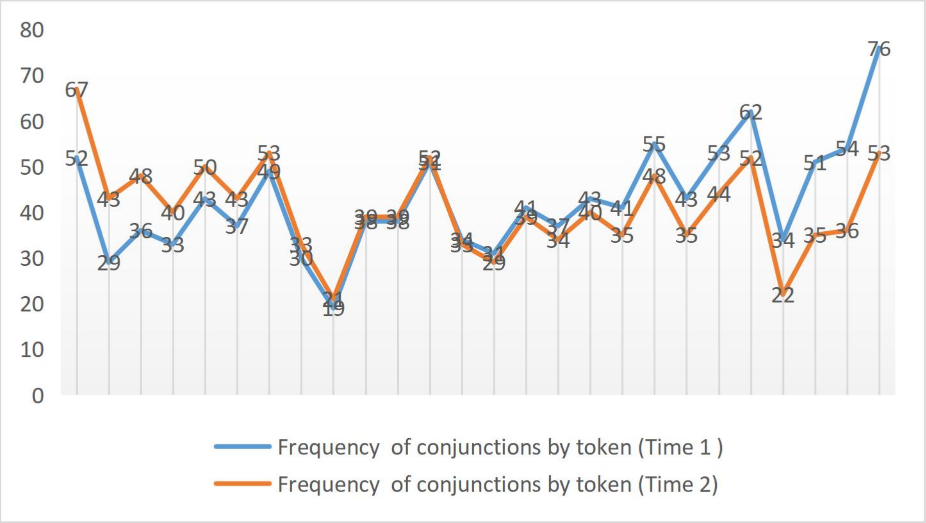


Chart 3. Difference in the frequency of conjunctions by type

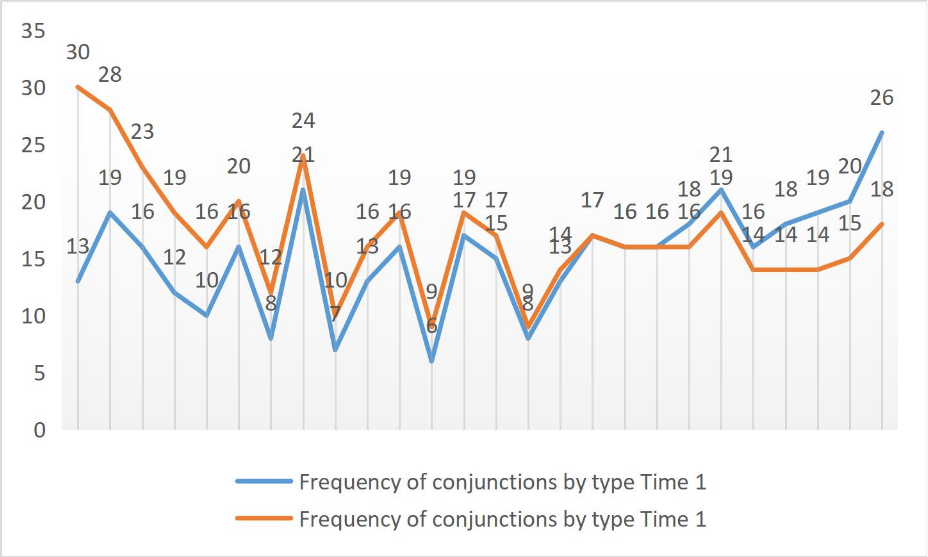


Table 11. Summary of the difference in the frequency of conjunctions across the two periods.

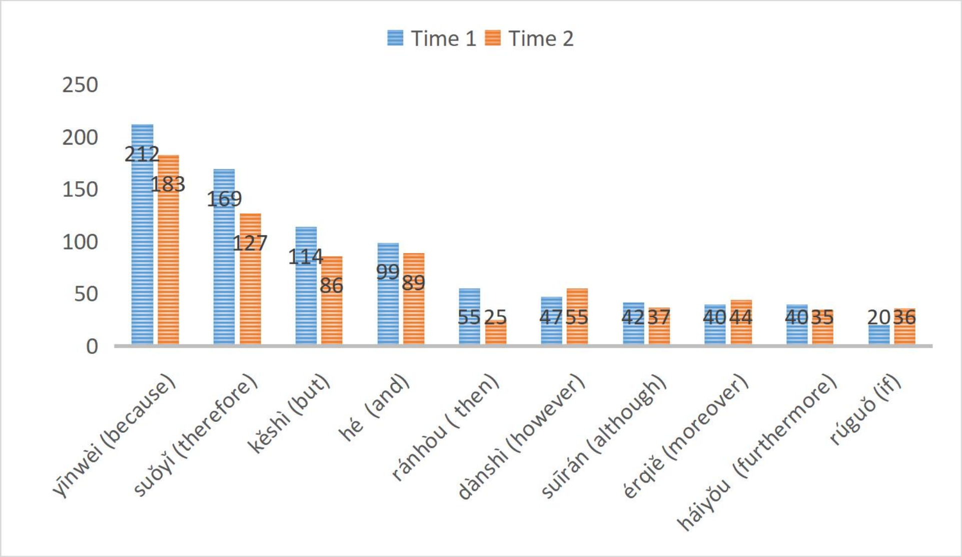
By token	Minimum	Maximum	Sum	Mean	Std. Deviation
Frequency of conjunctions Time1	19	76	1110	42.6923	11.97253
Frequency of conjunctions Time 2	21	67	1063	40.8846	10.25798
By type	Minimum	Maximum	Sum	Mean	Std. Deviation
Frequency of conjunctions Time1	6	26	397	15.2692	4.77123
Frequency of conjunctions Time 2	9	30	444	17.0769	5.09056

In Charts 2 and 3, there is no discernible overall trend of increase or decrease in frequency across the two periods, whether measured by type or by token. However, upon closer examination of individual development, it can be observed that among the 26 participants, when frequency is measured by token, only 12 participants used more conjunctions in Time 2 (approximately 46%), while the remaining 14 participants (around 54%) used fewer conjunctions during this period. On the other hand, when frequency is measured by type, 16 out of the 26 participants (over 60%) used more types of conjunctions in Time 2. Three participants remained the same (around 11%), and only 7 participants (around 29%) used fewer various conjunctions in Time 2. Additionally, 3 participants maintained the same amount of conjunction usage by type in both periods.

As depicted in Table 11, it's evident that, on average, participants used slightly fewer conjunctions in Time 2 when assessed by token. The mean quantity of conjunctions decreased from 42.69 in Time 1 to 40.88 in Time 2. In terms of total usage, the cumulative quantity of conjunctions used by all participants in Time 1, assessed by token, was 1,110, and in Time 2, it reduced to 1,063. However, a contrasting trend emerges when examining frequency by type, where its mean value increased from 15.27 to 17.07 across the two periods.The total amount also increased slightly from 397 to 444.

This comparison shows that, although there was no overall increase in total usage by token between the two periods, when examining the usage by type, it points to a noticeable rise in the diversity and total number of conjunctions used by participants in Time 2. This suggests that participants exhibited a tendency to employ a broader variety of conjunctions in Time 2.

Chart 4. Top 10 conjunctions used in Time 1 and Time 2



Upon examining the most frequently used conjunctions in two time periods, it can be seen that the top 10 conjunctions in both Time 1 and 2 are identical (the full list with Chinese can be found in appendix 3), differing only in their frequencies as illustrated in Chart 4. With the exception of 'dànshì,' 'érqiě,' and 'rúguǒ,' which see a slight increase in usage during Time 2, the remaining seven most frequently used conjunctions all demonstrate an downtick across the two periods. This observation suggests that participants, in Time 2, continued the pattern of favoring the use of common conjunctions, such as 'yīnwèi' and 'suǒyǐ,' albeit with a slightly lower frequency. The decreased frequency of most of the top-used conjunctions can be attributed to participants incorporating a greater variety of other conjunction types during Time 2.

In summary, the analysis of data indicates that the total usage of conjunctions by token remained relatively constant in Time 2, but there was an increase in the total amount by type. Despite a decrease in the frequency of the most commonly used conjunctions ('yīnwèi,' 'suǒyǐ,' etc.) in Time 2, participants continued to employ these conjunctions with lower frequency. This suggests that it is possible that more proficient L2 learners might have gained the ability to omit conjunctions when not obligatory. Furthermore, the observed rise in the variety of conjunctions in Time 2 is possibly attributed to two main factors. Firstly, it is possible that learners, now more proficient, deliberately chose to use a wider range of conjunctions. Secondly, the learning of four-character idioms contributed to this increase, as learners not only utilized common conjunctions but also experimented with new ones learned during this period. These learners may have acquired these conjunctions earlier but only began applying them appropriately in Time 2. Further elaboration on these findings will be provided in Chapter 6.

4.2.3 Frequency of V+N(object) high-frequency collocations

Different from the extraction of four-character idioms and conjunctions by CLAN, we obtained high-frequency V+N(object) high-frequency collocations from Sketch Engine using the uploaded corpus of diaries written by participants. We began by listing the verbs and then evaluated each one to check if it can form a collocation with a noun (object). To qualify as a collocation, a certain frequency needed to be attained. As specified in the Methodology section, we selected collocations with a Mutual Information (MI) score of ≥3. The MI scores of each verb with other elements in each sentence are all indicated within Sketch Engine. Subsequently, we compared the results with the native speaker corpus ZhTenTen from Sketch Engine to determine if the V+N(object) high-frequency collocations could also be identified in native speaker speech. This step aims to mitigate issues arising from the frequent use of peculiar collocations by L2 speakers. These procedures culminate in the compilation of the frequencies of V+N(object) high-frequency collocations used in Time 1 and 2, categorized by token in Chart 5 and by type in Chart 6.

Chart 5. Difference in V+N(object) high-frequency collocations frequency by token

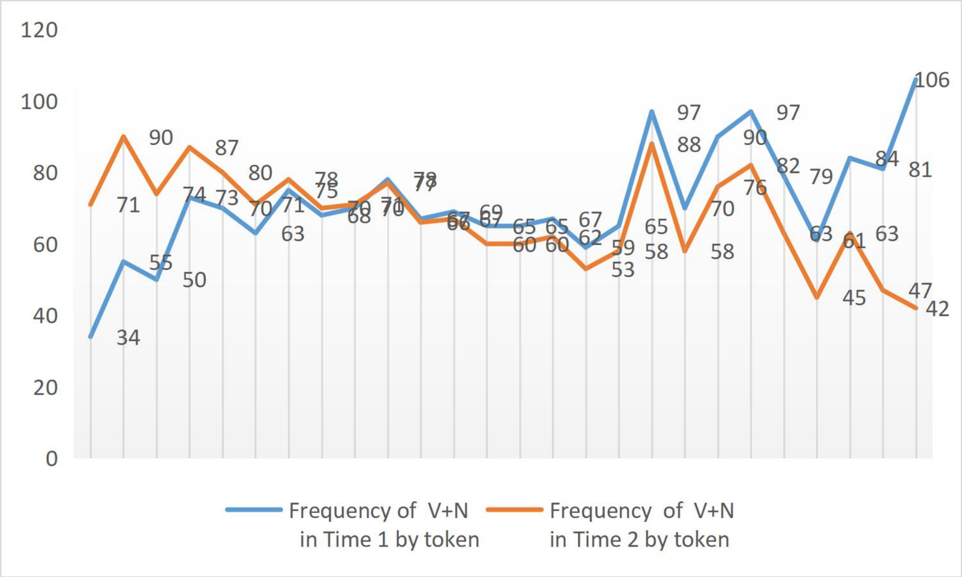


Chart 6. Difference in V+N(object) high-frequency collocations frequency by type

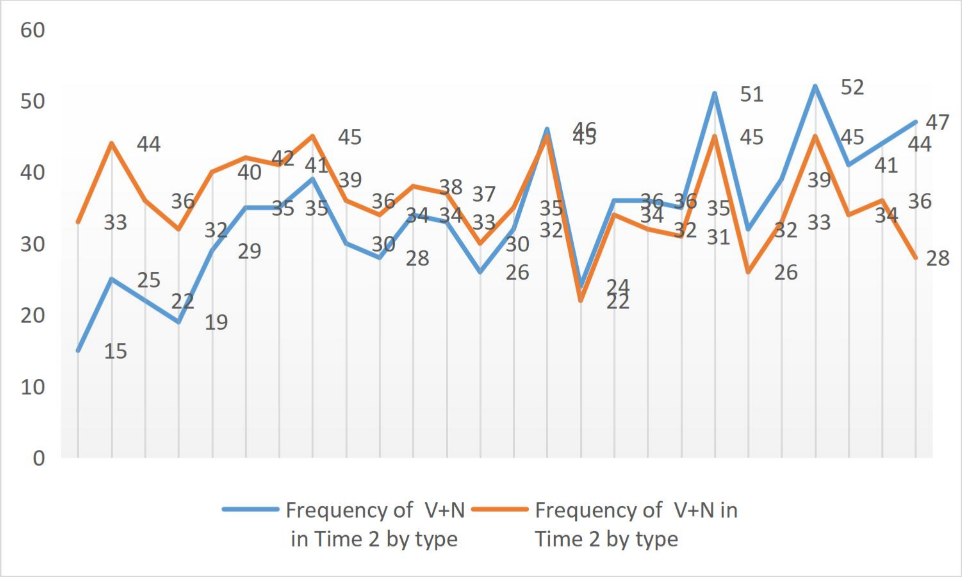


Table 12. Summary of V+N(object) high-frequency collocations frequency changes across the two periods.

By token	Minimum	Maximum	Sum	Mean	Std. Deviation
Frequency of V+N(object) high-frequency collocations **Time1**	34	106	1858	71.46	15.32
Frequency of V+N(object) high-frequency collocations **Time 2**	42	90	1759	67.65	12.94
By type	Minimum	Maximum	Sum	Mean	Std. Deviation

Frequency of V+N(object) high-frequency collocations **Time1**	15	52	885	34.04	9.37
Frequency of V+N(object) high-frequency collocations **Time 2**	22	45	934	35.92	6.18

Similar to the changes observed in the usage of conjunctions, the results regarding the use of V+N(object) high-frequency collocations by the participants are mixed; there is no clear overall decrease or increase evident from Chart 5 and 6. However, when examined individually, in Chart 5 measured by token, only 9 participants (35%) increased their use of V+N in Time 2. This figure rose to 54% when measured by type. It suggests that more participants tended to use a greater variety of V+N(object) high-frequency collocations in Time 2 compared to the quantity.

Another notable observation is that the frequency range of V+N(object) high-frequency collocations decreased in Time 2, regardless of whether measured by quantity or variety. Specifically, the minimum increased from 34 to 42 by token and from 15 to 22 by type, while the maximum decreased from 106 to 90 by token and from 52 to 45 by type. This indicates a reduction in the difference between the most and least capable participants of V+N, from 72 to 48 by token and from 37 to 23 by type.

When comparing the V+N(object) high-frequency collocations used in Time 1 and Time 2 by token, there is a slight decrease in the average usage (Table 12). However, if we compare the two periods by type, we observe an improvement from 34.03 to 35.92. Although this increase is not statistically significant, it indicates that participants tend to use a wider variety of V+N(object) high-frequency collocations in Time 2. Furthermore, both standard deviations and the range show a decrease when comparing groups by token and type, from 15.32 to 12.97 by token and from 9.37 to 6.18 by type and suggesting a more consistent usage in Time 2.

Table 13. The top 10 frequently used verbs in Time 1 and 2.

Ranking	Most frequent used verbs in **Time 1**	Frequency by Sketch Engine
1	qù '去' (go)	286
2	yǒu '有' (have)	267
3	chī '吃' (eat)	103
4	dào '到 ' (arrive)	100

5	xǐhuān '喜欢' (like)	79
6	kàn '看' (look)	78
7	xiǎng '想 ' (want/think)	73
8	shuō '说' (speak)	71
9	ràng '让' (let)	68
10	zuò '坐' (sit)	59
Ranking	Most frequent used verbs in **Time 2**	Frequency by Sketch Engine
1	qù '去' (go)	270
2	yǒu ' 有' (have)	245
3	shuō '说' (speak)	102
4	yào '要 ' (want)	87
5	ràng '让' (let)	75
6	kàn '看' (look)	72
7	chī '吃' (eat)	66
8	dào '到' (arrive)	66
9	xǐhuān '喜欢' (like)	59
10	lái '来' (come)	53

Table 14. Top 10 most frequently used verbs that formed V+N(object) high-frequency collocations in Time 1 and 2.

Ranking	Most frequent used verbs in **Time 1**	Frequency by Sketch Engine
1	yǒu 有 (have)	244
2	qù 去 (go)	191
3	chī 吃(eat)	80
4	zuò 坐(sit)	57
5	ràng 让(let)	50
6	dào 到(arrive)	48
7	méiyǒu 没有(don't have)	40
8	kàndào 看到 (see)	35
9	zuò做 (do)	35
10	lái 来 (come)	31
Ranking	Most frequent used verbs in **Time 2**	Frequency by Sketch Engine
1	yǒu 有(have)	230
2	qù去 (go)	151
3	ràng 让 (let)	56
4	chī吃 (eat)	40
5	méiyǒu 没有 (don't have)	32
6	shuō 说(speak)	23
7	mǎi 忙 (buy)	20
8	kàn 看 (look)	19

9	lái 来 (come)	18
10	zuò做 (do)	17

We extracted the top 10 most frequently used verbs and the verbs followed by noun (object)s that formed V+N(object) high-frequency collocations, as shown in Tables 13 and 14. Upon examining these tables, it becomes evident that 60% of the most frequently used verbs overlap with those forming V+N(object) high-frequency collocations. Specifically, in Time 1, the common verbs are yǒu (have), qù (go), chī (eat), zuò (do), ràng (let/make), dào (arrive). In Time 2, this percentage increased to 70%, including the verbs yǒu (have), qù (go), ràng (let/make), chī (eat), shuō (say), kàn (see), lái (come).This observation suggests that participants tended to maintain a preference for certain most frequently used verbs, similar to their preference in the usage of conjunctions, whether or not these verbs formed V+N(object) high-frequency collocations, across the two periods.

4.3 Evaluating the accuracy of formulaic sequences used in Time 1 and Time 2.

The investigation of the accuracy of four-character idioms, conjunctions, and V+N(object) high-frequency collocations in two periods aims to provide a detailed examination of the specific error patterns exhibited by participants and to determine whether there was an observable enhancement in their ability to employ formulaic sequences in Time 2.

4.3.1 Accuracy of four-character idiom

Four-character idioms possess relatively fixed forms and meanings compared to conjunctions and V+N(object) high-frequency collocations, inherently constrainting their contextual use. To assess participants' proficiency in using four-character idioms, it is crucial to evaluate their comprehension and ability to adapt these expressions within different contexts. Here, the context refers to their diaries, which frequently diverge from the classroom setting and possibly mirror real-life scenarios in the target country.

During Time 1, only two errors were identified in the use of four-character idioms, as illustrated in the following examples (full list as in Appendix 6). Each instance comprises the original Chinese text from the diaries which includes the target idiom marked in bold, its English translation, explanations of the idiom, along with indications of the errors and corrections, and a dictionary example. The explanations and dictionary samples of the idiom are primarily sourced from 'Chinese Idiom Dictionary' (2004) and the native speaker corpus *ZhTenTen* on Sketch Engine. Errors are denoted with an asterisk as shown in Example 1. For additional examples, please refer to Appendix 6 Example 1 to 7.

Example 1:

Diary: *有很多***十全十美**的花和美丽的水景.*

*'Yǒu hěnduō ***shíquánshíměi** de huā hé měilì de shuǐjǐng.'*

(There are many ***flawless** flowers and beautiful scenery with water.)

Explanation: 'shíquánshíměi' conveys a sense of extraordinary perfection and flawlessness. However, in this sentence, there is no specific emphasis on the flawlessness of the flowers but rather on their beauty. Therefore, a more suitable word could be simply 'beautiful'.

Dictionary good example:

*我们公司的产品都是**完美无缺**的，你可以放心使用。*

*'Wǒmen gōngsī de chǎnpǐn dōu shì **wánměi wúquē** de, nǐ kěyǐ fàngxīn shǐyòng.'*

(All products of our company are **flawless**; you can use them with confidence.)

In Time 1, only two errors (Example 1 and 2) errors occurred out of 21 idioms by token, while with the increase of frequency in Time 2, the errors also increase. There were 5 errors (Example 2 to 7) out of 43 idioms by token in Time 2. The correctness rate experienced a slight decrease from 90.48% to 88.37%. The 2 errors were made by the participant who applied most idioms (6) in Time 1, while the 5 errors were made by the two participants who both applied 6 idioms in Time 2.

Moreover, if we examine the errors one by one in both periods one by one, we can find out that all of them (2 errors in Time 1 and 5 errors in Time 2) are semantic errors, that is, the improper use of idioms in context.

4.3.2 Accuracy of conjunctions

To examine the accuracy of conjunctions, we assessed conjunctions both semantically and syntactically. We identified mainly three types of errors: discontinuous conjunctions with missing parts, incorrect combinations of conjunctions, and misuse of similar conjunctions. The first two types involve errors in both form and meaning, while the third one is related to semantic misuse.

Type 1: discontinuous conjunctions with missing parts (marked by '***_____**'). Discontinuous conjunctions are usually consisted of two parts. The participants used them with the first or second part missing, therefore the form and/or meaning was missing. These conjunctions include 除了...还 'chúle... hái' (besides... also);因为...所以 'yīn wèi... suǒ yǐ' (because... so);虽然...但(是) 'suī rán... dàn (shì)' (although... but); 如果...就 'rú guǒ... jiù' (if... then); 不仅...还 'bù jǐn... hái' (not only... but also).

Type 2: misuse of similar conjunctions.
We classified the conjunctions with similar meanings into three groups. The first group includes conjunctions that can express the meaning of 'and,'including *和 'hé'*(and); *还 'hái'* (also); *还有 'hái yǒu'*(furthermore); *而且 'érqiě'* (moreover),*并且 'bìngqiě'* (furthermore),*再加上'zài jiā shàng'* (in addition), illustrated by example 17,18 and 19. The second group comprise conjunctions which indicate the reason, including *由于 'yóuyú'* (due to); *为了 'wèile'* (in order to); *因为 'yīnwèi'* (because), explained by example 20 and 21 in Appendix 6.

Type 3: Incorrect combination of conjunctions.
For the discontinuous conjunctions, aside from the omission of one constituent, participants also demonstrated an inaccurate utilization of another. These conjunctions include 既然...那么 'jìrán...nàme' (since... then...); 即使...也 'jíshǐ...yě' (even if... still...); 要不然...会... 'yào bùrán...huì...'(otherwise... will...) which are presented from Example 22 to 25 in Appendix 6.

After examining each error, we compiled an accuracy list to assess whether any developments could be observed at Time 2. As illustrated in Table 15, during Time 1, the average correctness of conjunctions utilized in the diaries stood at 96.35%. This figure exhibited a marginal increase to 98.69% at Time 2. This high accuracy probably results from the avoidance, as observed in the usage of fixed-form formulaic sequences - the strategy that L2 learners use to avoid using expressions they feel comfortable with. Another possible reason is that, unlike four-character idioms, conjunctions are encountered more frequently by L2 learners, allowing them to master this type of formulaic sequence well through more exposure to input.

If we delve into each type of error, we can observe that Error Type 1, involving discontinuous conjunctions with missing parts, indicates that participants did not consider conjunctions as fixed forms with fillable elements in between. The more elements or interruptions between the two parts of conjunctions, the more errors participants made in using conjunctions. This represents an error in the form of formulaic sequences. Similarly, Error Type 2 involves incorrect combinations of conjunctions and also demonstrates the semantic misuse of conjunctions. As for Error Type 3, involving the incorrect combination of conjunctions, it once again reveals that participants were not aware that conjunctions are fixed form formulaic sequences, in which only the fillable parts can be flexible—the frame must remain fixed.

If we examine individual differences, we find that the minimum correctness increased from 75% to 88% between the two periods, accompanied by a notable decrease in the standard deviation (Table 15). This suggests that at Time 2, even the least proficient user of conjunctions displayed an enhanced proficiency in conjunction usage.

Table 15. Summary of accuracy of conjunctions differences in Time 1 and Time 2.

	Minimum	Maximum	Mean	Std. Deviation
Accuracy Time 1	75%	100%	96.35%	.05953
Accuracy Time 2	88%	100%	98.69%	.02753

4.3.3 Accuracy of V+N(object) high-frequency collocations

In addressing the accuracy of V+N(object) high-frequency collocations, as explained in the methodology section, Sketch Engine was employed instead of CLAN for a more lucid presentation of examples. Subsequent to the selection of V+N(object) high-frequency collocations, each instance was examined individually to ascertain its correct application. This examination revealed various several types of errors, encompassing incorrect collocation of verbs and nouns, misuse of serial verbs, the application of intransitive verbs as transitive verbs, instances of omission, and repetitions (from Example 26 to 37 in Appendix 6).

- **Error Type 1**: incorrect collocation of verbs and nouns, illustrated by Example 27, 28 and 29.
- **Error type 2**: error of usage of serial verbs.

In examples 30 to 33, we can observe that the participant has correctly used each main verb, such as *zǒu (go), kàn(watch), huí (return), tīng (hear)*. However, they have failed to select the correct directional or resultative verbs which formed 'serial verbs' in Chinese. For instance, in example 30, the correct form should be 'zǒuzhe qù' (walking-go) instead of 'zǒuguòqù' (walk-pass-go). In example 31, the serial verb should be replaced with a single verb, 'kàn' instead of 'kàndào.'In example 32, 'lái'(come) should be removed. In example 33, the result of 'tīng'(hear) should be 'chū,'(out) so it is 'tīng dé chū' (hear-able-out).

- **Error type 3: using intransitive verbs as transitive verbs**, illustrated by Example 34.
- **Error type 4: error of omission**, illustrated by Example 35 and 36.
- **Error type 5: error of repetition**, illustrated by Example 37 and 38.

By compiling all the errors in Time 1 and Time 2, we examined the change in correctness across the periods, as depicted in Table 16. Through statistical analysis, we observed that although the mean accuracy, assessed by token, showed a slight decrease from 96.39% to 95.9%, the minimum accuracy increased by around 3%.

Similar to the usage of conjunctions, the accuracy of V+N (object) high-frequency collocations is nearly at its peak (Table 16). One possible reason, aside from factors like avoidance and frequent encounters, is that the collocation V+N (object) often reflects the cognitive perception of L2 speakers, just as it does in their first language. For example, in English, we use "drink" to indicate the action of consuming liquids, such as water, milk, or juice. The same concept applies in Chinese. However, it's important to note that the mapping of verbs and concepts doesn't always align perfectly across languages, leading to the emergence of errors. Further details will be explored in Chapter 6.

Table 16. Summary of accuracy of V+N(object) high-frequency collocations changes in Time 1 and Time 2.

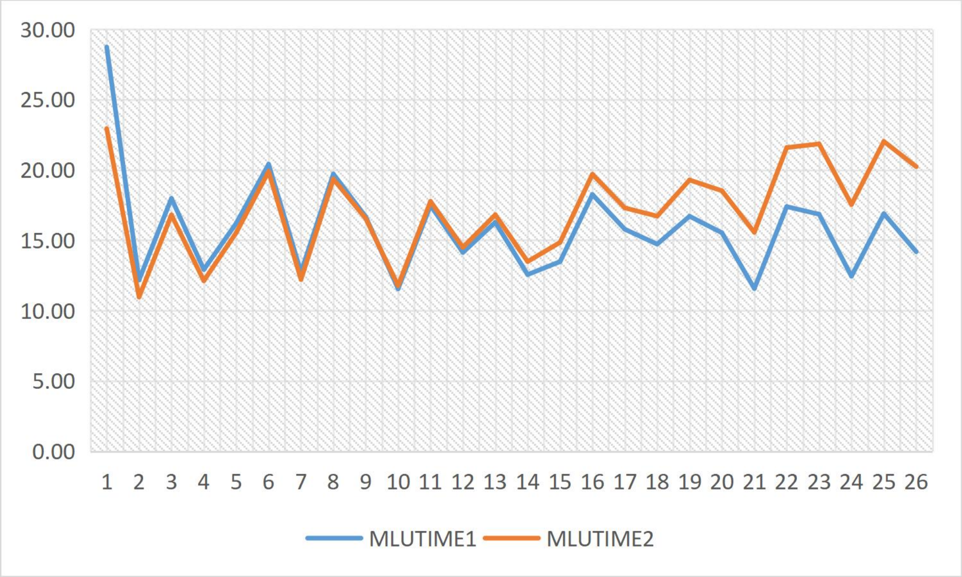
	Minimum	Maximum	Mean	Std. Deviation
Accuracy Time 1	85.71%	100%	96.39%	.0379210
Accuracy Time 2	88.89%	100%	95.90%	.0361815

4.4 Investigating the role of formulaic sequences in writing output across the periods measured by MLU and VOCD

4.4.1 Writing output measured by MLU

By running the MLU command, we extracted the mean sentence length of each set of diaries of each participant from CLAN and created Chart 6. However, it can be observed that one participant No.21 experienced the most significant decrease in MLU, going from 28.94 in Time 1 to 22.94 in Time 2, with a difference of 5.80. We examined the diaries of that participant (No.21) sentence by sentence and identified that this participant has some issues with punctuation usage, therefore when MLU is considered, we exclude this outlier when necessary to obtain a more reasonable understanding of the data.In Chart 6, if we exclude this outlier, it becomes evident that nearly half of the participants managed to increase MLU, while the remaining ones who did not manage to increase MLU at least maintained a similar level.

Chart 6. MLU of each set of diaries in Time 1 and Time 2.

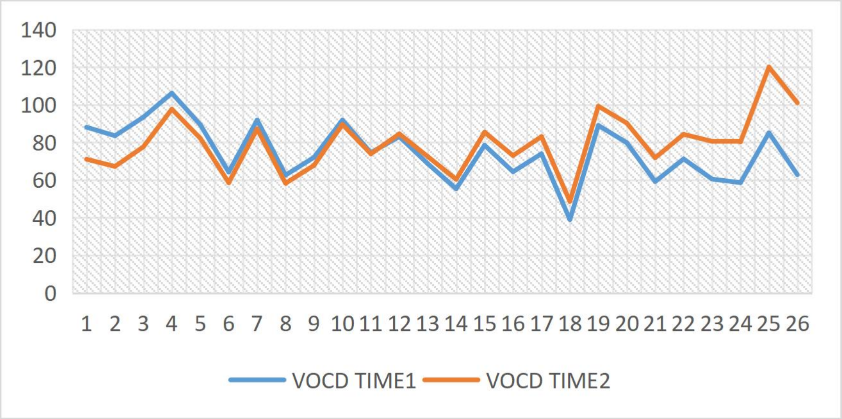


4.4.2 Writing output measured by lexical diversity

By running the VOCD command, we extracted the lexical diversity of each set of diaries from each participant using CLAN and created Chart 7.

Similar to the development in MLU, in Time 2, 15 out of 26 participants, which accounts for around 60%, experienced an increase in lexical diversity, with the maximum rise going from 63 to 101.

Chart 7. Lexical diversity of each set of diaries in Time 1 and Time 2.



4.4.3 Correlation between the use of four-character idiom and writing output in Time 1 and Time 2.

Running correlation analyses with the initial sample (N=26) by linear regression by SPSS, shown in Table 17, we found that the use of four-character idiom (by token) does not have a statistically significant impact on mean sentence length of the diaries (p > 0.05)). However, it has an association with lexical diversity in Time 1 (p ＜ .05), although not in Time 2 (p ＞ .05). The R Square value .162 indicates 16.2% of the variance in the dependent variable(lexical diversity) can be explained by the independent variable (four-character idioms). The correlation calculated using the frequency by type shows no significant difference from that using frequency by token.

Table 17. Correlation between four-character idiom frequency (by token)and MLU & lexical diversity

Coefficients

	Sig.	Sig.

	Time 1 (**R Square**)	Time 2 (**R Square**)
MLU[18]	p ＞ .05 (.018)	p ＞ .05 (.028)
Lexical Diversity[19]	**p ＜ .05 (.162)**	p ＞ .05 (.030)

4.4.4 Correlation between the use of conjunctions and writing output measured by MLU and lexical diversity in Time 1 and Time 2

Same analysis was carried out on the correlation between the use of conjunctions and MLU as well as lexical diversity, as shown Table 18. Only in Time 1 was there a correlation shown between the frequency of conjunctions used in diaries and lexical diversity(p＜ .05) with R-squared value of .176. It means that 17.6% of the variance in the dependent variable(lexical diversity) can be explained by the independent variable (conjunction frequency). However, from Table 19 generated from the linear regression analysis, it can be seen that the Unstandardised Coefficients value was negative, indicating a negative relationship. This implies that as participants use more conjunctions in their diaries, their lexical diversity will decrease. Regarding the link between the frequency of conjunctions and the mean sentence length, no statistical correlation can be seen(p＞ .05). The correlation calculated using the frequency by type shows no significant difference from that using frequency by token.

Table 18. Correlation between conjunction frequency[20] and MLU & lexical diversity

	Coefficients	
	Sig.	Sig.
	Time 1(**R Square**)	Time 2(**R Square**)
Mean Sentence Length	p ＞ .05 (.030)	p ＞ .05 (.084)
Lexical Diversity	**p ＜ .05 (.176)**	p ＞ .05 (.022)

Table 19. Unstandardised Coefficients of conjunction frequency and Lexical Diversity in Time 1

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std.	Beta		

[18] Independent^{a}

[19] Independent^{b}

[20] Dependent ^{2}

		Error			
Conjunction Frequency Time 1	**-0.331**	0.146	-0.419	-2.263	0.033

4.4.5 Correlation between V+N(object) high frequency collocation frequency and writing output in Time 1 and Time 2

The correlation between high-frequency V+N(object) collocation frequency and writing output, as measured by MLU and lexical diversity, differs when we measure frequency by token and type. When measured by token, no significant correlation (p > .05) was observed between V+N(object) high-frequency collocations and writing output, as shown in Table 20. However, when measuring frequency by type (Table 21), at Time 1, a correlation is observed between the dependent variable (lexical diversity) and the two independent variables (p < .05), with an R-squared value of .193. This indicates that 19.3% of the variance in the dependent variable can be explained by the independent variable (frequency by type).

Table 20. Correlation between V+N(object) high-frequency collocations frequency[21] (by token)and MLU & lexical diversity

	Coefficients	
	Sig.	**Sig.**
	Time 1(R Square)	**Time 2 (R Square)**
Mean Sentence Length	p > .05 (.001)	p > .05 (.000)
Lexical diversity	p > .05 (.001)	p > .05 (.004)

Table 21. Correlation between V+N(object) high-frequency collocations frequency[22] (by type) and MLU & lexical diversity

	Coefficients	
	Sig.	**Sig.**
	Time 1(R Square)	**Time 2 (R Square)**
Mean Sentence Length	p > .05 (.001)	p > .05 (.038)
Lexical diversity	**p < .05 (.193)**	p > .05 (.130)

[21] Dependent[3]

[22] Dependent[4]

4.5 Summary

In this chapter, our evaluation focuses on examining the frequency and accuracy of four-character idioms, conjunctions, and high-frequency V+N(object) collocations utilized in the diaries submitted in Time 1 and Time 2. Additionally, we conducted an analysis to ascertain the correlation between the usage of all three types of formulaic sequences and writing output, measured by mean sentence length and lexical diversity.

The results show that each of the three types of formulaic sequences exhibits a different trend. For fixed-form formulaic sequences like four-character idioms, the frequency almost doubled in total, while the accuracy remained high in both periods with a slight increase. For semi-fixed form formulaic sequences (conjunctions) and free-form formulaic sequences (V+N(object) high-frequency collocations), the frequency decreased by token and increased by type in Time 2. All three types of formulaic sequences show a correlation with lexical diversity in Time 1, although in the case of conjunctions, this correlation involves frequency only by type. In Time 2, such a correlation disappeared. As for MLU, no correlation was found in both Time 1 and Time 2.

Chapter 5. Results

Q1. What patterns can be seen in the usage of formulaic sequence according to their fixedness?

In examining the usage patterns of formulaic sequences according to their fixedness, we observed distinct characteristics for each type: four-character idioms with a fixed form, conjunctions with a semi-fixed form, and the most V+N(object) high-frequency collocations with a free form.

Fixed formulaic sequences:

- Frequency: The usage of fixed four-character idioms demonstrated the lost frequency of these three types, with only 21 in Time 1 and 43 in Time 2 by token. Moreover, not all participants employed them in their diaries.
- Accuracy: Despite the low frequency, accuracy remained relatively high, which is around 90% in Time 1 and around 88% in Time 2 (although still lower than the other two types). The errors were primarily attributed to semantic issues.

Semi-fixed formulaic sequences:

- Frequency: Semi-fixed conjunctions had a usage frequency between fixed four-character idioms and free-form collocations, which was around 1100 in both Time 1 and Time 2.
- Accuracy: conjunctions displayed a higher accuracy than fixed form sequence, which is around 96% in Time 1 and 98% in Time 2.The errors were caused by both semantic and form-related issues.

Free form formulaic sequences:

- Frequency: The most free-form V+N(object) high-frequency collocations are the most used formulaic sequences among the three types, which was empolyed around 1800 in both periods.
- Accuracy: Despite the highest frequency, accuracy of fixed form is almost as high as the semi-fixed formulaic sequences. The errors were caused by both semantic and form-related issues.

Q2: Was there any improvement of the usage of formulaic sequences, measured by frequency and accuracy, during the SA (from period 1 to period 2)?

When the usage of formulaic sequences was measured by frequency, we obtained the result that the fixed form type almost doubled in usage, while the semi-fixed and fixed types of formulaic sequences did not show an overall trend of increase; instead, significant individual differences can be seen. Regarding accuracy, on the other hand, almost all participants demonstrated a high level of mastery in all three types of formulaic sequences in both periods, which is nearly at the ceiling level.

Q3. Did Did the usage of the three types in of formulaic sequences, measured by frequency and accuracy, influence the lexical diversity and mean sentence length of the written texts respectively?

All three types of formulaic sequences exhibited a correlation with lexical diversity in Time 1, except in the case of conjunctions, where this correlation could be observed only when frequency was measured by type. However, by Time 2, this correlation disappeared. Regarding the correlation between the three types of formulaic sequences and MLU, no consistent correlation was found across both Time 1 and Time 2.

Chapter 6. Discussion

In the preceding chapter, I answered the research questions by drawing upon the discovered results. This chapter will further discuss pivotal aspects derived from these findings on each type of formulaic sequences, to further explore their usage, contributions and development.

6.1 Fixed form: four-character idioms

Frequency of four-character idioms

In the preceding chapters, it was observed that the three types of formulaic sequences are employed with varying frequencies. While all participants used conjunctions and V+N(object) high-frequency collocations, not everyone incorporated four-character idioms. Furthermore, for those who did include four-character idioms, the frequency was notably lower than that of the other two types.

One possible explanation for the diminished use of fixed four-character idioms compared to other types is their non-mandatory nature in writing, unlike verbs or function words such as conjunctions. Consequently, without specific instructions, the use of idioms is entirely at the discretion of L2 learners. Another contributing factor is the low semantic transparency of these idioms, where the meaning cannot be directly derived from their components. A comprehensive understanding of Chinese culture and literature is required to fully grasp and use these idioms in appropriate contexts.

On the other hand, the individual differences in frequency also indicate that the limited use of idioms may stem from L2 learners' tendency to write only what they are certain is correct, reflecting a form of avoidance strategy. Unlike speaking, where immediate responses are necessary, writing provides L2 learners with the time to choose alternative expressions for those they find challenging, such as the use of four-character idioms, effectively conveying their message in alternative ways (Laufer & Eliasson, 1993).

Accuracy

From the accuracy analysis, it is evident that no participant made mistakes on the form, such as writing only one part of the idiom and forgetting the rest. This observation, from one perspective, indicates that participants treat the idioms as a whole unit with a fixed form.

However, even the most fixed form formulaic sequences are, to some extent, flexible. For example, *飞扬跋扈* 'fēiyáng báhù', which means 'arrogant and domineering,' can be combined with other elements, as in the sentence *你飞扬也好，你跋扈也罢，这只能说明你的无能* ' Nǐ fēiyáng yě hǎo, nǐ báhù yě bà, zhè zhǐ néng shuōmíng nǐ de wúnéng' (Whether you are soaring high or acting arrogantly, it only demonstrates your incompetence) to satisfy the temporary expression needs (Wang & Gao, 2019). In the diaries submitted in two periods, no participant intended to break four-character idioms into pieces and tried to re-construct them, so no such flexibility of use was found. This practise, however, can be found in native speaker's output (Wang & Gao, 2019). It shows that these participants with an intermediate level of Chinese language, for the most fixed form formulaic sequence, or have not perceived that the most fixed form of formulaic sequences can also be inserted with other elements or they have not reached to the level of creating variations of fixed four-character idioms.

Learning four-character idioms in SA context

As mentioned in Chapter 4.2, despite the low frequency of four-character idioms, it is evident that, in Time 2, the total number of uses of four-character idioms more than doubled. This suggests that participants successfully incorporated more of these idioms into their diaries during the study abroad period. While it may be challenging to pinpoint the exact learning context for each idiom, whether in-class or out of class, contextual clues from the diaries provide some insights. For example, in a sentence extracted from a diary in Time 2, the use of the expression 'qū jí bì xiōng,' along with the practice of 'enter through the dragon's mouth and exit through the tiger's mouth,' indicates that these idioms were learned from local people.

6.2 Semi-fixed form: conjunctions

Awareness

Regarding conjunctions, as seen in examples 14 and 15, it is noticeable that mistakes occurred when more than one clause was inserted between 'yīn wèi... suǒ yǐ.' This may be attributed to the insertion of additional elements, making it easier for users to pay less attention to conjunctions, which function as frames for the sentences. If we envision conjunctions as the 'steel framework of a building to be filled with cement,' participants using only one part of conjunctions seemingly did not treat them as whole fixed frames that could be filled with other elements. Instead, they focused more on the meaning of the conjunctions themselves. Another contributing factor is that students are aware that sometimes conjunctions can be omitted, but they are not yet clear about the specific requirements for when such omissions are appropriate. This acutally could representstudents' initial attempts to flexibly apply formulaic sequences.

L1 transfer

For instance, in example 22, 'jìrán...nàme' was mistakenly used as 'jìrán...*suǒyǐ' (since....*so). 'jìrán...nàme' can be translated into English as 'since... then...'. However, in Chinese, both 'so' and 'then' can be used to express the result, but only 'nàme' can be used in conjunction with 'jìrán'.

Written Chinese VS Spoken Chinese

Another noteworthy observation pertains to the confusion between 'hái yǒu' (furthermore) and 'érqiě' (moreover), as exemplified in Example 18. While these terms share a similar meaning, they manifest slight differences. 'Hái yǒu' is more commonly used in spoken Chinese, and typically, participants should employ 'hái yǒu' less frequently than 'érqiě' in written contexts, even though diaries often exhibit a closer resemblance to spoken language compared to other forms, such as academic writing.

6.3 Free form: V+N(object) high frequency collocations

L1 transfer can also be observed in some errors of the misuse of free V+N(object) high frequency collocations. For instance, L2 learners used **rènshi gùshì* '*认识+故事' (*recognize + story). Both 'know' and 'recognize' can be translated into Chinese as *rènshi*, but *rènshi* is typically not used together with 'story' .

Over-generation

From examples 30 to 36, it is evident that one of the challenges participants faced is the use of serial verbs. Although participants demonstrated an awareness of the serial verb construction in the Chinese language, there was a tendency to overuse them. For instance, the verb 'dào' (arrive) was excessively used four times. However, in Time 2, this error decreased to only one occurrence. This reduction could be considered a potential sign of development in the participants' understanding and application of serial verbs over the course of the study.

6.4 Frequency of conjunctions and its influence on MLU

In Chapter 4 Data Analysis, we observed that the frequency of conjunctions have no correlation with the mean sentence length statistically. However, in some languages, conjunctions are essential for building longer sentences by joining clauses. Therefore, why did the usage of conjunctions did not show any influence to MLU? To understand the phenomenon, we first need to determine whether conjunctions are used frequently in Chinese in order to build longer sentences, similar to other languages such as English. We know that in English, if we intend to create a longer sentence, conjunctions can be employed as one of the ways to connect clauses. This is because English is hypotactic prominent, emphasizing the structure of language and requiring many connective words to ensure logical coherence and completeness in semantic meaning. Therefore, it is reasonable to expect that more connections lead to the formation of longer sentences in English. However, the situation is different in Chinese. Chinese discourses utilize fewer conjunctions than English, and their semantic relations are typically established through covert means. Consequently, Chinese requires fewer connectives, and semantic relations naturally evolve through the meanings of words or phrases (Yang, 2014). Therefore, the use of connections may not necessarily be linked to sentence length, as more proficient users can produce sentences with greater flexibility, incorporating a mixture of both long and short sentences, and may choose to use or omit conjunctions when necessary.

6.5 Summary

In Chapter 6, the study investigates the distinct characteristics and usage patterns of fixed form four-character idioms, semi-fixed expressions, and free form collocations among L2 Chinese learners. Notably, participants demonstrated varied frequencies in employing these formulaic sequences, with all consistently using conjunctions and high-frequency collocations, but not universally incorporating four-character idioms.

The discussion explores potential reasons for the lower utilization of fixed four-character idioms, including their non-mandatory nature and low semantic transparency. Additionally, the chapter sheds light on participants' accuracy in employing these idioms as cohesive units and their limited flexibility compared to native speakers. The study also examines the learning dynamics of four-character idioms in a study abroad context, revealing an increase in their usage over time. Furthermore, in the exploration of semi-fixed expressions, the chapter scrutinizes participants' awareness and occasional errors, emphasizing the evolving nature of their understanding and application of formulaic sequences. Finally, the study investigates the over-generation of free form collocations, observing challenges such as L1 transfer and instances of overuse. The chapter concludes with an examination of the frequency of conjunctions and their influence on mean sentence length, offering insights into the unique characteristics of conjunction usage in Chinese compared to languages like English.

Chapter 7 Conclusion

7.1 Implications of the study

Formulaic sequences play a pivotal role in language learning (Ellis, 2012). While previous research has predominantly emphasized the importance of formulaic sequences in spoken settings, there is a notable dearth of studies focused on their occurrence in written text. Moreover, given the growing emphasis on study abroad programs, understanding formulaic sequences in a study abroad context has become increasingly crucial. This study addresses this gap by collecting and analyzing diaries written by L2 Chinese participants during their study abroad experiences. The investigation encompasses formulaic sequences in Chinese across three distinct forms, ranging from the most fixed form of four-character idioms to semi-fixed form conjunctions and free form V+N(object) high-frequency collocations.

The frequency analysis conducted in this research examines the changes in the quantity of each type of formulaic sequence across two study abroad periods. The aim is to discern whether participants acquired more formulaic sequences or increased the variety in their usage. In tandem, the accuracy analysis zeroes in on identifying errors, aiming to uncover the underlying reasons and provide pedagogical insights for L2 Chinese formulaic sequence learning.

The implications of this thesis extend significantly to both second language learning and teaching. Firstly, the study offers valuable insights into errors observed in formulaic sequence usage and the associated challenges. This understanding is crucial for L2 Chinese learners and educators seeking to enhance language instruction, enabling them to tailor materials and strategies to address specific hurdles faced by learners. Additionally, the research sheds light on the influence of cultural and literary factors on formulaic sequence usage. Overall, the implications of this thesis transcend the specific linguistic context studied, providing valuable insights into effective language teaching methodologies and the nuances of L2 Chinese learners' linguistic development.

7.2 Contributions of the study

The multifaceted nature of formulaic sequences can pose challenges when attempting to comprehensively examine all their aspects. One of the key contributions of this study lies in shedding light on the examination of formulaic sequences from their most fixed form to their most free form. Their usage and the impact each type can have on writing output were investigated, aiming to identify the differences in the development of formulaic sequences, both in frequency and accuracy, during study abroad and the underlying possible reasons.

Few studies have been carried out to explore the inner differences between formulaic sequences in Chinese, focusing on their utilization, developments, and influences on writing output in the study abroad setting. This study aims to fill this gap through a thorough examination of all these factors.

Furthermore, employing diaries as a data resource also adds to the value of contributions. The diaries, with a relatively high reliability of authentication, offer valuable clues for tracing the origin of newly acquired lexicon during the learning process.

7.3 Limitations

There are several issues identified as limitations of the study. Due to the complexity of formulaic sequences, even though in this study, we chose from the most fixed type to the most free types and the ones in between, it is still far from including all the formulaic sequences used in our data resource, as formulaic sequences can cover more than half of our natural language.

Some other methodological issues are also observed in this study. The extraction of V+N(object) collocations, although assisted by Sketch Engine, is still primarily based on manual analysis. This highlights one of the most challenging aspects of formulaic sequences—the extraction, as a certain number of them are not continuous, making it difficult for software to extract without human intervention.

The same drawback applies to error analysis, which is carried out based on native speaker judgements. While native speakers can provide valuable insights into detecting errors, suggesting corrections, and summarising errors into types, it is still inherently subjective to some extent.

Regarding the corpus, a larger corpus would be helpful for a better understanding of the usage of formulaic sequences, especially the use of four-character idioms, which is challenging to comprehend when the participants do not employ them.

Moreover, the measurements of writing output are limited to MLU and lexical diversity. Their correlation with formulaic sequences was examined through a purely quantitative approach. However, this approach may not capture certain improvements made by L2 learners, such as readability and native likeness, which are challenging to evaluate quantitatively.

7.4 Suggestions for future research

Due to the limitations inherent in measuring writing output, it is essential to consider additional tools for evaluation, including those designed for assessing readability and native-likeness. The advancement of Natural Language Processing (NLP) and artificial intelligence provides an opportunity for cross-disciplinary research aimed at investigating L2 learning.

References

Abdel Latif, M. M. (2012). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics*, *34*(1), 99–105. https://doi.org/10.1093/applin/ams073

Arnon, I., & Christiansen, M. H. (2017). The role of Multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science*, *9*(3), 621–636. https://doi.org/10.1111/tops.12271

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing*, *18*(3), 191–208. https://doi.org/10.1016/j.jslw.2009.05.003

Bardovi-Harlig, K. (2002). A new starting point? *Studies in Second Language Acquisition*, *24*(2), 189–198. https://doi.org/10.1017/s0272263102002036

Beals, D. E. (2002). The CHILDES project: Tools for Analyzing Talk: Vol. 1. transcription format and programs; vol. 2. the database (3rd ed.). B. MacWhinney. Mahwah, NJ: Erlbaum, 2000. pp. 366 (vol. 1); pp. 418 (vol. 2). *Applied Psycholinguistics*, *23*(2), 304–306. https://doi.org/10.1017/s0142716402222079

Becker, J. D. (1975). The phrasal lexicon. *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing - TINLAP '75*. https://doi.org/10.3115/980190.980212

Bruton, D. L., & Kirby, D. R. (1987). Research in the classroom: Written fluency: Didn't we do that last year? *The English Journal*, *76*(7), 89. https://doi.org/10.2307/818661

Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, *6*(2). https://doi.org/10.5070/l462005216

Che, W., Feng, Y., Qin, L., & Liu, T. (2021). N-LTP: An open-source neural language technology platform for Chinese. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. https://doi.org/10.18653/v1/2021.emnlp-demo.6

Chenowth, N. A., & Hayes, J. R. (2001). Fluency in writing. *Written Communication*, *18*(1), 80–98. https://doi.org/10.1177/0741088301018001004

Cicourel, A. V. (2007). A personal, retrospective view of ecological validity. *Text &amp; Talk*, *27*(5–6). https://doi.org/10.1515/text.2007.033

Conti, S. (2017). Chengyu in Chinese Language Teaching: A Preliminary Analysis of Italian Learners' Data. *Irish Journal of Asian Studies*, *3*(1), 59–82.

Cordier, C. (2013). *The presence, nature and role of formulaic sequences in English advanced learners of French: A longitudinal study*. University of Newcastle upon Tyne.

Coxhead, A. & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16, 129-147. doi:10.1016/j.jslw.2007.07.002.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (2014). An automated approach to the phrasicon of EFL learners. *Learner English on Computer*, 67–79. https://doi.org/10.4324/9781315841342-5

Du, H. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal*, *97*(1), 131–143. https://doi.org/10.1111/j.1540-4781.2013.01434.x

Duran, P. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220–242. https://doi.org/10.1093/applin/25.2.220

Dörnyei, Z., Durow, V., & Zahran, K. (2004). Individual differences and their effects on formulaic sequence acquisition. *Language Learning &amp; Language Teaching*, 87–106. https://doi.org/10.1075/lllt.9.06dor

El-Dakhs, D. A. S., Prue, T. T., & Ijaz, A. (2017). The effect of the explicit instruction of formulaic sequences in pre-writing vocabulary activities on foreign language writing. *International Journal of Applied Linguistics and English Literature*, 6(4), 21-31.

Ellis, N. C. (1996). Sequencing in SLA. *Studies in Second Language Acquisition*, *18*(1), 91–126. https://doi.org/10.1017/s0272263100014698

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44. https://doi.org/10.1017/s0267190512000025

Ellis, N. C., & Sinclair, S. G. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 234–250. https://doi.org/10.1080/713755604

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text - Interdisciplinary Journal for the Study of Discourse*, *20*(1). https://doi.org/10.1515/text.1.2000.20.1.29

Freed, B. F. (1995). *Second language acquisition in a study abroad context*. Benjamins.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, *26*(02). https://doi.org/10.1017/s0272263104262064

Godfrey, L., Treacy, C., & Tarone, E. (2014). Change in French second language writing in study abroad and domestic contexts. *Foreign Language Annals*, *47*(1), 48–65. https://doi.org/10.1111/flan.12072

Golafshani, N. (2015). Understanding reliability and validity in qualitative research. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2003.1870

Granger, S. (1998). Prefabricated patterns in Advanced efl writing: Collocations and formulae. *Phraseology*, 145–160. https://doi.org/10.1093/oso/9780198294252.003.007

Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield, & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 21e33). Basingstoke, UK: Palgrave Macmillan.

He, X. (2023). Effects of structural complexity and L1 experience on L2 acquisition of Chinese multiword sequences. *Foreign Language Annals*, *56*(2), 480–500. https://doi.org/10.1111/flan.12678

Housen, A. E., Vedder, I. E., & Kuiken, F. E. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy andfluency in SLA. Language Learning & Language Teaching. volume 32*. John Benjamins Publishing Company.

Howarth, P. (1998). Phraseology and Second language proficiency. *Applied Linguistics*, *19*(1), 24–44. https://doi.org/10.1093/applin/19.1.24

Hunston, S. (2002). *Corpora in Applied Linguistics*. https://doi.org/10.1017/cbo9781139524773

Ife, A., Vives Boix, G., & Meara, P. (2000). The impact of study abroad on the vocabulary development of different proficiency groups. *Spanish Applied Linguistics*, *4*(1), 55–84.

Johnson, M. D., Mercado, L., & Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing*, *21*(3), 264–282. https://doi.org/10.1016/j.jslw.2012.05.011

Kecskes, I., & Sun, C. (2017). *Key Issues in Chinese as a Second Language Research*. https://doi.org/10.4324/9781315660264

Khoualdi, S. (2017). Raising teachers' awareness of the significance of formulaic sequences in writing proficiency. *Human Sciences Journal*, 79-93.

Kilgarriff, A., & Kosem, I. (2012). Corpus tools for lexicographers. *Electronic Lexicography*, 31–56. https://doi.org/10.1093/acprof:oso/9780199654864.003.0003

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten Years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kinginger, C. (2008). Language learning in study abroad: Case studies of Americans in France. *The Modern Language Journal*, *92*(s1), 1–124. https://doi.org/10.1111/j.1540-4781.2008.00821.x

Kinginger, C. (2009). *Language Learning and Study Abroad*. https://doi.org/10.1057/9780230240766

Kinginger, C. (2011). Enhancing language learning in study abroad. *Annual Review of Applied Linguistics*, *31*, 58–73. https://doi.org/10.1017/s0267190511000031

Knoch, U. (2007). *Diagnostic writing assessment: the development and validation of a rating scale* (thesis). The University of Auckland, New Zealand.

Laufer, B., & Eliasson, S. (1993). What causes avoidance in L2 learning. *Studies in Second Language Acquisition*, *15*(1), 35–48. https://doi.org/10.1017/s0272263100011657

Leow, R. P. (1993). Fluency and accuracy: Toward balance in language teaching and learning. Hector Hammerly. Clevedon, UK: Multilingual matters, 1991. pp. VIII + 208. $79.00 cloth, $27.00 paper. *Studies in Second Language Acquisition*, *15*(2), 267–269. https://doi.org/10.1017/s0272263100012055

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*(2), 85–102. https://doi.org/10.1016/j.jslw.2009.02.001

MacWhinney, B. (2000). The CHILDES project: Tools for Analyzing Talk (Third Edition): Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics*, *26*(4), 657–657. https://doi.org/10.1162/coli.2000.26.4.657

Manchón, R. M., & Polio, C. (2021). L2 writing and language learning. *The Routledge Handbook of Second Language Acquisition and Writing*, 1–6. https://doi.org/10.4324/9780429199691-1

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

Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2. *Studies in Second Language Acquisition*, *21*(1), 49–80. https://doi.org/10.1017/s0272263199001023

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

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pérez-Vidal, C., & Barquin, E. (2014). Chapter 9. comparing progress in academic writing after formal instruction and study abroad. *AILA Applied Linguistics Series*, 217–234. https://doi.org/10.1075/aals.13.11ch9

Reynolds, D. W. (2005). Linguistic correlates of Second language literacy development: Evidence from middle-grade learner essays. *Journal of Second Language Writing*, *14*(1), 19–45. https://doi.org/10.1016/j.jslw.2004.09.001

Sanz, C., & Morales-Front, A. (2018). *The Routledge Handbook of Study Abroad Research and Practice*. https://doi.org/10.4324/9781315639970

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing*, *9*(3), 259–291. https://doi.org/10.1016/s1060-3743(00)00028-x

Sasaki, M. (2004). A multiple-data analysis of the 3.5-year development of EFL student writers. *Language Learning*, *54*(3), 525–582. https://doi.org/10.1111/j.0023-8333.2004.00264.x

Savignon, S. J. (2017). Communicative competence. *The TESOL Encyclopedia of English Language Teaching*, 1–7. https://doi.org/10.1002/9781118784235.eelt0047

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing, and use*. John Benjamins.

Serrano, R., Tragant, E., & Llanes, À. (2012). A longitudinal analysis of the effects of one year abroad. *The Canadian Modern Language Review*, *68*(2), 138–163. https://doi.org/10.3138/cmlr.68.2.138

Sinclair, J. (1999). *Corpus, concordance, Collocation*. Oxford University Press.

Siyanova-Chanturia, A. (2015). *Collocation in beginner learner writing: A longitudinal study. System*, 53, 148-160.

Snellings, P., Van Gelderen, A., & De Glopper, K. (2004). The effect of enhanced lexical retrieval on Second language writing: A classroom experiment. *Applied Psycholinguistics*, *25*(2), 175–200. https://doi.org/10.1017/s0142716404001092

Spelman Miller, K. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research*, *4*(2), 123–148. https://doi.org/10.1191/136216800675510135

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, *18*(2), 103–118. https://doi.org/10.1016/j.jslw.2009.02.003

Taguchi, N. (2008). Building language blocks in L2 Japanese: Chunk learning and the development of complexity and fluency in spoken production. *Foreign Language Annals*, *41*(1), 132–156. https://doi.org/10.1111/j.1944-9720.2008.tb03283.x

Taguchi, N. (2011). The effect of L2 proficiency and study-abroad experience on pragmatic comprehension. *Language Learning*, *61*(3), 904–939. https://doi.org/10.1111/j.1467-9922.2011.00633.x

Taguchi, N., Li, S., & Xiao, F. (2013). Production of formulaic expressions in L2 Chinese: A developmental investigation in a study abroad context. *Chinese as a Second Language Research*, *2*(1), 23–58. https://doi.org/10.1515/caslar-2013-0021

Waibel, B. (2008). *Phrasal verbs: German and Italian learners of English compared*. Saarbrucken, Germany: VDM.

Wang, S. (2020). *Chinese Multiword Expressions*. https://doi.org/10.1007/978-981-13-8510-0

Wolfe-Quintero, K. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i Press.

Wood, D. C. (2015). *Fundamentals of formulaic language: An introduction*. Bloomsbury Academic, An imprint of Bloomsbury Publishing Plc.

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

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

Wright, C. (2019). Developing communicative competence in adult beginner learners of Chinese. *The Routledge Handbook of Chinese Language Teaching*, 134–148. https://doi.org/10.4324/9781315104652-9

Wright, C. (2020). Effects of task type on L2 Mandarin fluency development. *Journal of Second Language Studies*, *3*(2), 157–179. https://doi.org/10.1075/jsls.00010.wri

Wright, C., & Schartner, A. (2013). 'I can't … I won't?' international students at the threshold of Social Interaction. *Journal of Research in International Education*, *12*(2), 113–128. https://doi.org/10.1177/1475240913491055

Xue, N., & Shen, L. (2003). Chinese word segmentation as LMR tagging. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing* -. https://doi.org/10.3115/1119250.1119278

Yuldashev, A., Fernandez, J., & Thorne, S. L. (2013). Second language learners' contiguous and discontiguous multi-word unit use over time. *The Modern Language Journal*, *97*(S1), 31–45. https://doi.org/10.1111/j.1540-4781.2012.01420.x

钱旭菁 (2008). *汉语语块研究初探*,《北京大学学报：哲学社会科学版》, 第 5 期, 139-146.

王文龙. (2017). *国际汉语初级阶段语块构建研究*, 北京大浮出版社

孙梦梅, 于金淳. (2000)《*汉语成语字典*》, 商务印书馆国际有限公司

张斌, 范开泰, 张亚军. (2000). *现代汉语语法分析*, 华东师范大学出版社

周健. (2007). *语块在对外汉语教学中的价值与作用*,《暨南学报：哲学社会科学版》, 第 1 期, 99-104.

张斌, 范开泰, 张亚军. (2000). *现代汉语语法分析*, 华东师范大学出版社.

Appendix

Appendix 1 Four-character idioms in Time 1 and Time 2

Four-character idioms in Time 1			
Chinese	**pīnyīn**	**Translation**	**Frequency**
大开眼界	dà kāi yǎn jiè	Expand one's horizons	2
久而久之	jiǔ ér jiǔ zhī	Over a long period of time	1
力所不及	lì suǒ bù jí	Beyond one's capability	1
一丝不苟	yī sī bù gǒu	Meticulous; with utmost care	1
一五一十	yī wǔ yī shí	Accurate and detailed	1
一干二净	yī gān èr jìng	Thorough and clean; neat and tidy	1
十全十美	shí quán shí měi	Perfect in every aspect	1
兴致勃勃	xìng zhì bó bó	With great interest and enthusiasm	1
叹为观止	tàn wéi guān zhǐ	To marvel at something; to find something amazing	1
无话可说	wú huà kě shuō	Speechless; have nothing to say	1
激动人心	jī dòng rén xīn	Exciting and stirring	1
迫不及待	pò bù jí dài	Unable to wait; impatient	1
无所事事	wú suǒ shì shì	Idle; doing nothing	1
手忙脚乱	shǒu máng jiǎo luàn	Busy and disorderly; in a mess	1
受宠若惊	shòu chǒng ruò jīng	Overwhelmed by favor; feeling flattered	1
流连忘返	liú lián wàng fǎn	To be so captivated that one forgets to leave	1
大开眼界	dà kāi yǎn jiè	Expand one's horizons	1
喜气洋洋	xǐ qì yáng yáng	Full of joy and happiness	1
闻所未闻	wén suǒ wèi wén	Hear something one has never heard before	1
人山人海	rén shān rén hǎi	Crowded with people; a sea of people	1
感恩不尽	gǎn ēn bù jìn	Feel grateful beyond words	1
Four-character idioms in Time 2			
Chinese	**pīnyīn**	**Translation**	**Frequency**
大开眼界	dà kāi yǎn jiè	Expand one's horizons	4
人山人海	rén shān rén hǎi	Crowded with people; a sea of people	3
独一无二	dú yī wú èr	Unique; one of a kind	2
流连忘返	liú lián wàng fǎn	To be so captivated that one forgets to leave	2
提心吊胆	tí xīn diào dǎn	Apprehensive and fearful	2
总而言之	zǒng ér yán zhī	In short; to sum up	2
远近闻名	yuǎn jìn wén míng	Well-known far and near	1
趋吉避凶	qū jí bì xiōng	Seeking good fortune and avoiding disaster	1

谢天谢地	xiè tiān xiè dì	Thanking the heavens and the earth; expressing gratitude	1
谈虎色变	tán hǔ sè biàn	Changing color at the mention of a tiger; terrified at the thought of a tiger	1
虎头蛇尾	hǔ tóu shé wěi	Starting energetically but ending weakly	1
筋疲力尽	jīn pí lì jìn	Exhausted physically and mentally	1
笨手笨脚	bèn shǒu bèn jiǎo	Clumsy; awkward	1
焉知非福	yān zhī fēi fú	How do you know it's not a blessing? (Every cloud has a silver lining)	1
毫无疑问	háo wú yí wèn	Without a doubt	1
有备而来	yǒu bèi ér lái	Come prepared; be ready for anything	1
时移俗易	shí yí sú yì	Times change and customs evolve	1
数不胜数	shù bù shèng shù	Too numerous to count; countless	1
截然不同	jié rán bù tóng	Completely different	1
惊涛骇浪	jīng tāo hài làng	Turbulent waves; stormy seas	1
心满意足	xīn mǎn yì zú	Content and satisfied	1
尽力而为	jìn lì ér wéi	Do one's best; make every effort	1
家常便饭	jiā cháng biàn fàn	Commonplace; everyday occurrence	1
好吃懒做	hào chī lǎn zuò	Enjoy eating and be lazy; indulge in pleasure and avoid work	1
大汗淋漓	dà hàn lín lì	Sweating profusely	1
大同小异	dà tóng xiǎo yì	More or less the same; similar with minor differences	1
塞翁失马	sài wēng shī mǎ	Misfortune may be a blessing in disguise	1
吞吞吐吐	tūn tūn tǔ tǔ	Mumbling and hesitating in speech	1
吃喝玩乐	chī hē wán lè	Eating, drinking, playing, and having fun	1
千差万别	qiān chā wàn bié	Vastly different; a thousand differences	1
刻骨铭心	kè gǔ míng xīn	Deeply engraved in one's memory	1
利大于弊	lì dà yú bì	The benefits outweigh the disadvantages	1
乱七八糟	luàn qī bā zāo	In a mess; chaotic	1
久而久之	jiǔ ér jiǔ zhī	Over a long period of time	1

Appendix 2 Most frequent used conjunctions in Time 1 and Time 2

Chinese	**pīnyīn**	**Translation**
因为	yīnwèi	because
所以	suǒyǐ	therefore
可是	kěshì	but
和	hé	and
然后	ránhòu	then
但是	dànshì	however
虽然	suīrán	although
而且	érqiě	moreover
还有	háiyǒu	furthermore

如果	rúguǒ	if

Appendix 3 Most frequent used verbs in Time 1 and Time 2

Most frequent used verbs in Time 1			
Ranking	**Chinese**	**pīnyīn**	**Translation**
1	去	qù	go
2	有	yǒu	have
3	吃	chī	eat
4	到	dào	arrive
5	喜欢	xǐhuān	like
6	看	kàn	look
7	想	xiǎng	want/think
8	说	shuō	speak
9	让	ràng	let
10	坐	zuò	sit
Most frequent used verbs in Time 2			
Ranking	**Chinese**	**pīnyīn**	**Translation**
1	去	qù	go
2	有	yǒu	have
3	说	shuō	speak
4	要	yào	want
5	让	ràng	let
6	看	kàn	look
7	吃	chī	eat
8	到	dào	arrive
9	喜欢	xǐhuān	like
10	来	lái	come

Appendix 4. Frequency of each type of formulaic sequence used in diaries in Time 1&2 by token.

No.	Four-character idioms Time 1	Four-character idioms Time 2	Conjunctions Time 1	Conjunctions Time 2	V+N Time 1	V+N Time 2	MLU in Time 1	MLU Time 2	VOCD Time 1	VOCD Time 2
1	3	1	34	33	70	80	16.922	16.92	16.63	87.17
2	6	4	37	43	90	76	14.065	14.07	15.94	82.21
3	0	2	38	39	69	67	16.465	16.47	16.07	97.67
4	1	0	33	40	68	70	13.879	13.88	14.27	77.68
5	1	4	53	44	79	63	12.218	12.22	13.30	72.48
6	0	1	43	50	70	58	16.413	16.41	21.15	71.07

7	1	2	38	39	65	58	12	12.00	10.78	85.41
8	1	0	37	34	84	63	11.403	11.40	14.82	67.82
9	2	2	31	29	97	88	15.544	15.54	14.80	90.27
10	0	0	51	35	50	74	12.411	12.41	11.75	73.94
11	0	7	49	53	65	60	19.946	19.95	19.20	71.83
12	1	2	43	35	78	77	17.457	17.46	23.10	89.53
13	1	5	34	22	63	71	14.707	14.71	53.40	99.24
14	0	0	36	48	55	90	15.229	15.23	38.15	67.22
15	0	0	29	43	70	71	26	26.00	26.03	101.05
16	0	0	41	39	67	66	27.148	27.15	21.50	84.31
17	0	1	54	36	106	42	16.035	16.04	24.85	60.38
18	0	0	62	52	65	60	15.383	15.38	16.50	80.24
19	0	0	41	35	59	53	16.92	16.92	17.51	58.58
20	1	4	55	48	81	47	11.294	11.29	11.67	83.10
21	1	0	43	40	61	45	30.304	30.30	23.42	120.02
22	0	0	52	67	75	78	14.935	14.94	17.57	80.57
23	0	0	51	52	67	62	17.391	17.39	15.64	48.65
24	1	3	30	33	73	87	13.469	13.47	19.69	84.53
25	0	0	19	21	34	71	12.5	12.50	11.60	58.28
26	0	2	76	53	97	82	17.689	17.69	19.04	72.98

Appendix 5. Frequency of each type of formulaic sequence used in diaries in Time 1&2 by type.

No.	Four-character idioms Time 1	Four-character idioms Time 2	Conjunctions Time 1	Conjunctions Time 2	V+N Time 1	V+N Time 2	MLU Time 1	MLU Time 2	VOCD Time 1	VOCD Time 2
1	3	1	16	19	46	45	16.92	16.63	91.89	87.17
2	6	4	12	19	44	36	14.07	15.94	89.43	82.21
3	0	2	17	19	52	45	16.47	16.07	106.19	97.67
4	1	0	16	16	34	38	13.88	14.27	93.42	77.68
5	1	4	17	17	47	28	12.22	13.30	68.70	72.48
6	0	1	13	14	32	35	16.41	21.15	88.08	71.07
7	1	2	13	16	36	34	12.00	10.78	78.49	85.41
8	1	0	8	12	32	26	11.40	14.82	72.07	67.82
9	2	2	16	14	41	34	15.54	14.80	79.85	90.27
10	0	0	18	16	22	36	12.41	11.75	74.53	73.94
11	0	7	19	28	30	36	19.95	19.20	59.19	71.83
12	1	2	21	19	36	32	17.46	23.10	91.87	89.53
13	1	5	20	15	35	42	14.71	53.40	89.03	99.24
14	0	0	10	16	25	44	15.23	38.15	83.53	67.22
15	0	0	13	30	29	40	26.00	26.03	62.84	101.05
16	0	0	16	16	33	37	27.15	21.50	71.22	84.31

17	0	1	19	14	39	33	16.04	24.85	55.35	60.38
18	0	0	16	20	26	30	15.38	16.50	58.61	80.24
19	0	0	7	10	24	22	16.92	17.51	64.23	58.58
20	1	4	18	14	35	31	11.29	11.67	73.96	83.10
21	1	0	16	23	28	34	30.30	23.42	85.11	120.02
22	0	0	21	24	39	45	14.94	17.57	60.51	80.57
23	0	0	8	9	19	32	17.39	15.64	38.97	48.65
24	1	3	15	17	35	41	13.47	19.69	83.10	84.53
25	0	0	6	9	15	33	12.50	11.60	62.61	58.28
26	0	2	26	18	51	45	17.69	19.04	64.41	72.98

Appendix 6. Examples of errors of all three types of formulaic sequences in diaries

No.	Details
Example 1	Diary: 有很多*十全十美的花和美丽的水景. 'Yǒu hěnduō *shíquánshíměi de huā hé měilì de shuǐjǐng.' (There are many *flawless flowers and beautiful scenery with water.) Explanation: 'shíquánshíměi' conveys a sense of extraordinary perfection and flawlessness. However, in this sentence, there is no specific emphasis on the flawlessness of the flowers but rather on their beauty. Therefore, a more suitable word could be simply 'beautiful'. Dictionary good example: 我们公司的产品都是完美无缺的，你可以放心使用。 'Wǒmen gōngsī de chǎnpǐn dōu shì wánměi wúquē de, nǐ kěyǐ fàngxīn shǐyòng.' (All products of our company are flawless; you can use them with confidence.)
Example 2	Diary: 看到这么漂亮的塔,我们都*无可话说。 'Kàn dào zhème piàoliang de tǎ, wǒmen dōu *wúkě huà shuō.' (Seeing such a beautiful tower, we all *have nothing to say.) Explanation: 'wúkě huà shuō' means 'having nothing to say'. It is obvious that the expression that the participant intended to use was to say that 'we are unable to express our feelings in words,' therefore the correct phrase should be 无法用言语表达 'Wúfǎ yòng yányǔ biǎodá'. Sketch Engine good example: 我实在无话可说, 因为我已经出离愤怒了。 'Wǒ shízài wúhuàkěshuō, yīnwèi wǒ yǐjīng chūlí fènnù le.' (I truly have nothing to say because I'm already beyond angry.
Example 3	Diary:我为何说的很*吞吞吐吐？ 'Wǒ wèihé shuō de hěn *tūntūntǔtǔ?' (Why do I speak so *hesitantly?) Explanation: In this sentence, given the context, the participant intended to express their lack of fluency in Chinese. Therefore, the correct phrase should be 'kēkēbābā (haltingly)' instead of 'tūntūntǔtǔ.' 'Tūntūntǔtǔ' conveys the idea of speaking with hesitation and reluctance. Dictionary good example:

	你有话直说, 不必吞吞吐吐。 'Nǐ yǒu huà zhí shuō, bù bì tūntūntǔtǔ.' (Speak your mind; there's no need to be hesitant.)
Example 4	Diary: 即日我决定会趁著*惊涛骇浪去冲浪。 'Jírì wǒ juédìng huì chéngzhe *jīngtāohàilàng qù chōnglàng.' (Today, I've decided to go surfing amidst the *terrifying waves.) Explanation: The term 'jīngtāohàilàng' means 'terrifying and surging waves', commonly used metaphorically for perilous environments or experiences. The accurate expression that the participant intended to refer to was the presence of larger waves, so it could be replaced by 浪大 'làngdà' (big waves). Dictionary good example: 在惊 涛 骇 浪 中, 我 们 的 船 几 次 面临被吞噬的危险。 'Zài jīngtāohàilàng zhōng, wǒmen de chuán jǐ cì miànlín bèi tūnshì de wēixiǎn.' (In the midst of the terrifying waves, our boat faced the danger of being swallowed several times.)
Example 5	Diary:英国跟台湾有*数不胜数的差别。 'Yīngguó gēn Táiwān yǒu *shùbùshèngshù de chābié.' (There are *countless differences between the UK and Taiwan.) Explanation: The expression 'shùbùshèngshù' can be translated into English as 'countless' or 'innumerable.' It signifies an inability to count or an extremely large quantity, describing a magnitude that surpasses ordinary counting or enumeration capabilities. This idiom is commonly used to depict a vast number of things or phenomena that are beyond regular quantification. However, a more suitable expression to describe the difference between Taiwan and the UK in this context would be shùbùguòlái '数不过来' (too numerous to count). Dictionary good example: 在 我 们学 校 里, 好 人 好事 层 出 不穷, 数不胜数。 'Zài wǒmen xuéxiào lǐ, hào rén hàoshì céng chū bù qióng, shùbùshèngshù.' (In our school, there are countless kind people and good deeds emerging constantly.)
Example 6	Diary:因为游泳是我唯一一个目标，所以我没*有备而来了。 'Yīnwèi yóuyǒng shì wǒ wéiyī yīgè mùbiāo, suǒyǐ wǒ méi *yǒu bèi ér lái le.' (Because swimming is not my forte, I came un-*prepared for a goal-oriented approach.) Explanation:The idiom 'yǒubèi ér lái' can be translated into English as 'come prepared.' It means arriving or engaging in something with thorough preparation, emphasizing readiness for various situations in a specific activity, task, or situation. However, in the given context, the use of this idiom may not be entirely appropriate, as it suggests a lack of alternative solutions rather than a lack of thorough preparation. Sketch Engine good example: 此次 展会 上 , 针对 海外 市场有备而来的 展品 往往 很 受 欢迎 。 'Cǐcì zhǎnhuì shàng, zhēnduì hǎiwài shìchǎng yǒubèiérlái de zhǎnpǐn wǎngwǎng hěn shòu huānyíng.' (At this exhibition, products prepared for the overseas market are often well-received.)
Example 7	Diary:下坡我很紧张因为我怀疑刹车的成效。我*提心吊胆会突然断。 'Xiàpō wǒ hěn jǐnzhāng yīnwèi wǒ huáiyì shāchē de chéngxiào. Wǒ *tíxīndiàodǎn huì túrán duàn.' (Downhill, I feel nervous because I doubt the effectiveness of the brakes. I'm *on edge, fearing they might suddenly fail.) Explanation:The expression 'tíxīndiàodǎn' conveys a continuous state of worry without specifying details. It accurately describes a persistent sense of anxiety where both heart and courage are suspended, creating a vivid image of extreme fear and apprehension. Unlike other examples, the semantic meaning of the idiom was appropriate for the context. However, the subject of what the participant is feared of is missing.

	Dictionary good example:看他的杂技 表演, 真让 人 提心吊胆, 怕 他摔下来。 'Kàn tā de zájì biǎoyǎn, zhēn ràng rén tíxīndiàodǎn, pà tā shuāi xiàlái.' (Watching his acrobatic performance really makes people on the edge afraid he might fall.) According to the identified errors above, it is evident that in Time 1, two errors occurred out of 21 idioms by token. In Time 2, there were 5 errors out of 43 idioms by token. Consequently, the correctness rate experienced a slight decrease from 90.48% to 88.37%. Notably, most of these errors were attributed to the misuse of the semantic meaning of idioms rather than incorrect forms.
Example 8	chúle... hái' (apart from... also) 除了打渔，租船*　让我们探索当地的海岸，很有意思。 Chúle dǎ yú, zū chuán *　ràng wǒmen tànsuǒ dāngdì de hǎi'àn, hěn yǒu yìsi. (Apart from fishing, renting a boat also allows us to explore the local coastline, which is very interesting.) 'Apart from' has two usages: the first one means 'excluding', in which 'also' can be omitted, as in the sentence 其他人都去了，除了你 'qítā rén dōu qù le, chúle nǐ'(Everyone went, except you). The second meaning is 'in addition to this'. In this case, as example 8 shows, 'also' must be added.
Example 9	虽然我很兴奋，*____我紧张极了。 Suīrán wǒ hěn xīngfèn, *____wǒ jǐnzhāng jíle. (Although I am very excited, *____ I am extremely nervous.)
Example 10	虽然网上的广告说四张床 *____ 只有两张。 Suīrán wǎngshàng de guǎnggào shuō sì zhāng chuáng *____ zhǐyǒu liǎng zhāng. (Although the online advertisement claiming four beds, *____ there are only two.)
Example 11	虽然他们吃苦，*　他们不愿意放弃这个工作，因为他们的情形都非常严重。 Suīrán tāmen chī kǔ, *____tāmen bù yuànyì fàngqì zhège gōngzuò, yīnwèi tāmen de qíngkuàng dōu fēicháng yángròu. (Although they endure hardships, *____ they are unwilling to give up this job because their situations are very serious.)
Example 12	和其他日记不同，*____这篇日记我写的很快 1*。但是这样能让我有机会说我的假期计划。 Hé qítā rìjì bùtóng, *____ zhè piān rìjì wǒ xiě de hěn kuài*。2 Dànshì zhèyàng néng ràng wǒ yǒu jīhuì shuō wǒ de jiàqī jìhuà. (Unlike other diaries, *____ I wrote this diary entry quickly. However, doing so gives me the opportunity to talk about my holiday plans.) *Note 1: the punctuation here should be comma instead of full stop.
Example 13	rú guǒ... jiù' (if... then) 所以如果我有空时间的话，我*____去沙滩静静地休息一下。 Suǒyǐ rúguǒ wǒ yǒu kòng shíjiān dehuà, wǒ *____ qù shātān jìngjìng de xiūxí yīxià. (So if I have some free time, I *____ go to the beach to relax quietly.)
Example 14	yīn wèi... suǒ yǐ' (because... so): 因为我对花生过敏，还网上看的文章都说大部分的台湾菜都有花生或花生油，*_____我感觉脆弱。 Yīnwèi wǒ duì huāshēng guòmǐn, hái wǎngshàng kàn de wénzhāng dōu shuō dà bùfèn de Táiwān cài dōu yǒu huāshēng huò huāshēng yóu, *_____ wǒ gǎnjué cuìruò. (Because I am allergic to peanuts, and articles I read online also say that most Taiwanese dishes contain peanuts or peanut oil, *_____I feel vulnerable.)
Example 15	yīn wèi... suǒ yǐ' (because... so): 因为我晕车，而且最近及格了台湾的驾驶考试，*_____决定租车开车去日月潭。 Yīnwèi wǒ yùnchē, érqiě zuìjìn jígéle Táiwān de jiàshǐ kǎoshì, *_____ juédìng zūchē kāichē qù Rìyuè Tán. (Because I get motion sickness, and also, I recently passed the driving test in Taiwan, *_____decided to rent a car and drive to Sun Moon Lake.)
Example 16	'bù jǐn... hái' (not only... but also) 但是我不仅想家。我*_____特别想我的男朋友，我不能过一天不跟她讲电话。 Dànshì wǒ bùjǐn xiǎng jiā. Wǒ *_____tèbié xiǎng wǒ de nán péngyǒu, wǒ bù néng guò yī tiān bù gēn tā jiǎng diànhuà.

	But I not only miss home. I *_____ miss my boyfriend a lot. I can't go a day without talking to him on the phone.
Example 17	hé'(and) and 'bìngqiě' (furthermore) 我们先参观鹅銮鼻灯塔和鹅銮鼻公园，接着去了旁边的沙滩休息*和晒太阳浴 。 'wǒmen xiān cānguān Éluánbí dēngtǎ hé Éluánbí Gōngyuán, jiēzhe qùle pángbiān de shātān xiūxí *hé shài tàiyángyù.' (We first visited the Eluanbi Lighthouse and Eluanbi Park, then went to the nearby beach to rest *and sunbathe.) When indicating two actions happening simultaneously or sequentially, 'bìngqiě' should be used instead of 'hé'. Here, 'resting on the beach' and 'sunbathing' are two actions that occur sequentially so 'hé' is incorrect.
Example 18	hái yǒu'(furthermore) and 'érqiě' (moreover) 风景漂亮极了*还有人很少。 'Fēngjǐng piàoliang jíle, * hái yǒu rén hěn shǎo.' (The scenery is extremely beautiful, *and there are very few people.) 'hái yǒu' indicates parallel, complementary, or progressive relationships. 'érqiě' refers to the continuity of existence, indicating something additional or alongside the mentioned item. Here, 'scenery is beautiful' and 'few people' are in a parallel relationship, complementing each other. Therefore, 'érqiě' should be used.
Example 19	zàijiāshàng'(in addition) and 'érqiě'(moreover) 台湾有无数旅游目的地，*再加上各地传统的食物千差万别，比如说扬子江的饺子、台南的芒果冰，海岸的蚵仔煎。 Táiwān yǒu wúshù lǚyóu mùdìdì, *zàijiāshàng gèdì chuántǒng shíwù qiānchāwànbie, bǐrú shuō Yángzǐ Jiāng de jiǎozi, Táinán de mángguǒ bīng, hǎi'àn de háizǐ jiān (Taiwan has numerous tourist destinations, and, *in addition, traditional foods from various regions are incredibly diverse, such as dumplings from the Yangtze River, mango ice from Tainan, and oyster omelets from the coast.) In this sentence, 'numerous tourist destinations in Taiwan', and 'the cuisine is also diverse' two aspects are in a parallel relationship to describe Taiwan, so 'érqiě' should be used to indicate this connection instead of 'zàijiāshàng' which means 'in addition.'
Example 20	yóuyú' (due to) and 'yīnwèi' (because) *由于她明亮的眼睛，这个细节跟爸爸一样特别令人难忘。 'Yóuyú tā míngliàng de yǎnjīng, zhège xìjié gēn bàba yīyàng tèbié lìngrén nánwàng.' (Due to her bright eyes, this detail is as uniquely memorable as my dad.) 'Yóuyú' is commonly accompanied by a full sentence, conveying a thorough explanation or a cause-and-effect relationship, whereas 'yīnwèi' allows for a more free usage. Since 'her bright eyes' is a noun phrase, it is better suited for a flexible usage similar to 'yīn wéi'.
Example 21	wèile' (in order to) and 'yīnwèi' (because) Bond 在对付敌人表现很大的克制，*为了他想把完庭放在第一位。 'Bánde zài duìfù dírén biǎoxiàn hěn dà de kèzhì, wèile tā xiǎng bǎ wántíng fàng zài dì yī wèi' (Bond shows a great deal of restraint in dealing with enemies, *in order to he wants to prioritize loyalty.) 'Wèile' means 'in order to' and is employed to convey purpose. In this context , the reason for Bond's restraint is family-related; therefore, 'yīnwèi' (because) should be used.
Example 22	jìrán...nàme' (since... then...) 既然花我买不到，*所以我决定为朋友买一个凤梨，因为凤梨是最像花的。 'jìrán huā wǒ mǎi bù dào, *suǒyǐ wǒ juédìng wèi péngyǒu mǎi yīgè fènglí, yīnwèi fènglí shì zuì xiàng huā de' (Since I can't buy flowers, *so I decided to buy a pineapple for my friend, because a pineapple is the most flower-like.)
Example 23	jìrán...nàme' (since... then...) 既然是台湾网站, 我不能用英国的信用卡，*所以我需要给旅馆打电话解释我的问题....。 'Jìrán shì Táiwān wǎngzhàn, wǒ bùnéng yòng Yīngguó de xìnyòngkǎ, *suǒyǐ wǒ xūyào gěi lǚguǎn dǎ diànhuà jiěshì wǒ de wèntí.' (Since it's a Taiwanese website, I can't use my British credit card, *so I need to call the hotel to explain my situation....)
Example	jíshǐ...yě' (even if... still...)

24	*既然我想我家人，我也常常感觉孤独。 '*jìrán wǒ xiǎng wǒ jiārén, wǒ yě chángcháng gǎnjué gūdú' (*Since I miss my family, I still often feel lonely.)
Example 25	yào bùrán...huì...'(otherwise... will...) 但是，对我来说我觉得最好我不要带太多的东西，要不然我的行李太重了，*所以搬不动。 'Dànshì, duì wǒ lái shuō wǒ juédé zuì hǎo wǒ bùyào dài tài duō de dōngxi, yàobùrán wǒ de xínglǐ tài zhòng le, *suǒyǐ bān bù dòng.' (However, for me, I think it's best not to bring too much stuff; otherwise, my luggage will be too heavy, *so I am not able to move it.)
Example 26	...一个朋友给我*打短信，叫我和另外一个朋友去天文馆。 '...yī gè péngyǒu gěi wǒ *dǎ duǎnxìn, jiào wǒ hé lìngwài yī gè péngyǒu qù tiānwén guǎn.' (...a friend *called me by text , asking me to go to the observatory with another friend.) Explanation: To send a text in Chinese, the verb 发' fā' (send) should be used instead of 打 'dǎ', which means 'call'.
Example 27	...也给他们看政大大学的员工*寄给我的短信。 '......yě gěi tāmen kàn Zhèng Dà Dàxué de yuángōng *jì gěi wǒ de duǎnxìn.' (I also showed them the text messages *delivered to me by the employees of NCCU University.)
Example 28	*参观 了 两 个 很 有趣 的 校外 考察 旅行 。 '*Cānguān le liǎng gè hěn yǒuqù de xiàowài kǎochá lǚxíng.' (*Visited two very interesting off-campus study trips.) Explanation: The verb 'visit' is not suitable in this context because people typically participate in or take trips, rather than visiting them.
Example 29	可是从来连一次都没自己做过*餐”。 'Kěshì cónglái lián yīcì dōu méi zìjǐ zuò guò *cān.' (But I have never cooked a *meal by myself) The verb 'cook' should be used with 饭 'fàn') which means meal as well, just as cān.
Example 30	...因此 我*走过去 一 个 酒吧。 '.....yīncǐ wǒ *zǒu guòqù yī gè jiǔbā' (Therefore, I * walked over to a bar.)
Example 31	...一起去*看到电影。 'Let's go * watch-arrive a movie together. ' (.....yīqǐ qù *kàndào diànyǐng.)
Example 32	在星期五我们*回来台北 。 'zài xīngqīwǔ wǒmen *huílái Táiběi.' (On Friday, we * return-come to Taipei.)
Example 33	你往往*听到同学们的国家。 'nǐ wǎngwǎng *tīngdào tóngxuémen de guójiā.' (You often * hear-arrive classmates' countries.)
Example 34	而且我担心六个月会不会不够*进步我的中文。 'Érqiě wǒ dānxīn liù gè yuè huì bù huì bù gòu *jìnbù wǒ de zhōngwén.' (And I'm worried that six months may not be enough to *improve my Chinese.) 'Jìnbù' (improve) is an intransitive verb, which means it typically does not take a direct (object). In this context, it should be replaced with 'tígāo,' which is a transitive verb with the meaning 'enhance' or 'improve.'
Example 35	“......喜*___ 咖喱 饺子” 'xǐ*___kālí jiǎozi' (.....enjoy*___curry dumplings.) In this sentence, 'huān' (joy) is missing, which together with 'xǐ'(like) to form the verb 'xǐhuān' (like). Same error can be found in example 36, in which the correct verb should be 达到'dádào'(arrive-reach).
Example 36	“虽然真难达[*]那个水平” Suīrán zhēn nán dá [*] nèi ge shuǐpíng. 'Although it's really hard to reach [*] that level.
Example 37	......*操说流利的中文。 '......*cāo shuō liúlì de Zhōngwén.'

	'......* speak speak fluent Chinese.' In this example, only one verb is needed. It can be chose from or cāo (speak) or shuō (speak) when followed by language as (object).