The Development of Relative Clauses in L2 Chinese: A Corpus-Based Study

Li-ping Chang

Abstract Acquisition of relative clauses (RCs) in a second language (L2) has long been a popular research focus, particularly in determining whether L2 learners' acquisition of RCs conforms to the Noun Phrase Accessibility Hierarchy (NPAH) (Keenan & Comrie, 1977), which proposes that subject-extracted RCs are the easiest to learn because they are the most commonly produced RC type with the fewest error rate. Early studies have mostly focused on Indo-European languages, especially English. In this study, we adopt a corpus-based approach to analyze the distribution of subject-extracted RCs (SRCs) and object-extracted RCs (ORCs) by Chinese learners with six different L1s and at two proficiency levels to test whether SRCs are easier than ORCs for Chinese L2 learners. The corpus we used is the Test of Chinese as a Foreign Language (TOCFL) Learner Corpus comprised 4,709 compositions written by test-takers of the writing section. A total of 2,055 RCs are analyzed, including 1,362 RCs at the CEFR-B1 (intermediate-high) level and 693 RCs at the CEFR-B2 (advanced) level by native speakers of English, Spanish, Japanese, Korean, Vietnamese, and Indonesian, representing three different language typologies. From the perspectives of RCs occurring in the grammatical position in the matrix sentence and the animacy of the head noun, the results show that ORCs for Chinese L2 learners are easier than SRCs. These results go against the NPAH hypothesis. In addition, no matter what branching types (i.e. left, right, or left-and-right) the learner's native language was, all lower-proficiency level language learners produced more ORCs than SRCs. These results coincide with the development pattern of RCs for L1 Chinese acquisition. Therefore, we propose that the dominant factor in learning Chinese RCs is word order, since ORCs have the same SVO word order as Chinese simple sentences. Regardless of learners' language background, learners can produce ORCs more naturally and with more ease. After the L2 language proficiency increases, SRCs will take over that advantage and learners' language use will become gradually closer to the target language.

L. Chang (✉)
Graduate Program of Teaching Chinese As a Second Language, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei 106, Taiwan
e-mail: lchang@ntu.edu.tw

H. H.-J. Chen et al. (eds.), *Learner Corpora: Construction and Explorations in Chinese and Related Languages*, Chinese Language Learning Sciences,
https://doi.org/10.1007/978-981-19-5731-4_7

Keywords Relative clause · L2 acquisition · Interlanguage · Learner corpus · Mandarin Chinese

1 Introduction

Ever since Keenan and Comrie (1977) proposed their Noun Phrase Accessibility Hierarchy (NPAH), the relative clause (RC) has received special attention from linguists and language acquisition researchers. Using the linguistic typology approach, Keenan and Comrie conducted a thorough survey of over 50 different languages and proposed the following hierarchy of relativized noun phrases: SU>DO>IO>OBL>GEN>OCOMP. This supposes that the easiest noun phrase to relativize in a sentence is the subject, followed by the direct object, indirect object, prepositional noun, subordinate noun clause, and the object of a comparative sentence. Scholars later applied the NPAH to language acquisition research. Many previous studies indicate that, as learners acquire each type of RC, the difficulty order conforms to the NPAH. That is, the subject-extracted RC (SRC) is the easiest to be acquired (Doughty, 1991; Eckman et al., 1988; Gass, 1979; Izumi, 2003). However, the results predicted by the NPAH apply mainly to Indo-European languages, particularly English. In these languages, the RC follows the head noun that it modifies, as in example (1a) where the head noun 'person' occurs at the left of the RC. This structure is the opposite in Chinese, as in (1b) where the head noun *ren* 'person' occurs at the right.

(1) a. the person who bought the book
b. *mai shu de ren*
buy book DE person
'the person who bought the book'

The question of whether or not languages placing the head noun at the right (e.g. Japanese or Mandarin Chinese) also conforms to the NPAH prediction that has been the subject of inquiry for decades. There are still disputes over the results of that research. Tarallo and Myhill's (1983) cross-language research indicates that, for Japanese or Chinese, the object-extracted RC (ORC) is easier than the SRC. Hasegawa (2005) also supports this result for the Japanese. However, Sakamoto and Kubota (2000) studied learners whose native language was English, Chinese, or Indonesian and found that they all conformed to NPAH. Regarding L2 Chinese RC learning, there are mixed results either supporting NPAH (Li, 2015; Xu, 2014) or rejecting it (Dai, 2010; Tarallo & Myhill, 1983).

Many previous studies have been conducted with cognitive experiments, for example, by combining two sentences into a single sentence or judging the grammaticality of RCs. Those experiments were conducted within a controlled environment. The advantage of such experiments is that a feature effect can be pinpointed and focused, but only a very limited number of samples can be observed, thus perhaps causing incomplete and disputable results. In this research, we turn to naturally

produced various interlanguages instead of using a limited sample produced in a controlled way. In order to deal with such a large amount of influencing features, we adopted a corpus-based approach using the Test of Chinese as a Foreign Language (TOCFL) Learners' Corpus comprising 2,259 compositions written by learners of different proficiency levels and various L1s, including English, Japanese, Korean, Vietnamese, Indonesian, and Spanish, which provides naturally composed data for our analysis. The goal of our research is to uncover the patterns of RC acquisition of L2 Chinese through a corpus approach. These are the three research questions we ask:

1. What is the distribution between SRCs and ORCs produced by L2 learners?
2. Is there any different distribution among learners of different L1 backgrounds?
3. Is there any difference? If so, what is the difference?

In order to present the results and uncover influencing factors, we provide a comprehensive review of the related issues.

2 Related Research on Chinese RC Acquisition

2.1 Disputed Results on the Acquisition of SRCs and ORCs

As mentioned in the previous section, results that conform to the NPAH prediction apply mainly to Indo-European languages, the head nouns of which occur at the left of RCs. The question of whether left-branching languages (e.g. Japanese, Korean, and Mandarin Chinese) also conform to the NAPH prediction has led to increased research within the last decade. Sakamoto and Kubota (2000) investigated the RC acquisition of Japanese L2 learners with different L1s (i.e. English, Mandarin Chinese, and Indonesian) by using a sentence-combining task. The results conformed to the NAPH prediction. However, later studies on L2 learners of Japanese did not completely conform to the NAPH prediction (Ozeki & Shirai, 2007). Their results suggest that the NPAH does not predict the difficulty order of Japanese RCs. See Hasegawa (2005) for further reading. O'Grady et al. (2003) explored the second language acquisition of Korean. 53 native English speakers were asked to select a corresponding picture based on the type of RCs they heard. Participants showed that ORCs were more difficult to comprehend than SRCs. These results conform to the NAPH prediction. Tarallo and Myhill's (1983) cross-language research indicates that, for English native speakers learning languages where the RC occurs after a head noun (e.g. German, Portuguese, and Persian), the SRC proves easier than the ORC. This appears to conform to the NAPH prediction, but they also found that the reverse occurs when English native speakers are learning Japanese or Chinese. In such cases, the ORC is easier than the SRC.

Regarding RC acquisition for L2 Chinese learners, Packard's (2008) research utilizes a self-paced reading task to assess English speakers' processing difficulty

of L2 Chinese RCs. The results show that English speakers demonstrate slower processing times for SRCs. Packard suggests that Chinese instructors should teach ORCs before SRCs. Since this study targets English-speaking participants, this suggestion may only be applicable to English native speakers. A different approach used by Dai (2010) also supports that ORCs are easier than SRCs. He employed a sentence-combining task to investigate how the position and type of RC impact Chinese language acquisition. His 39 participants (intermediate to advanced level proficiency) came from various L1 backgrounds (i.e. English, Japanese, and Korean). That study concluded that the RC position has no significant effect on the acquisition of RCs. However, the type of RC has an obvious impact on learners' acquisition. The order of acquisition indicates that ORCs are the easiest, followed by SRCs, ID-RCs, OBL-RCs, etc. Therefore, the acquisition of L2 Chinese RCs does not support the NPAH hypothesis.

Despite those results, there are other studies which suggest that the NPAH acquisition theory is applicable to L2 Chinese. For example, Xu (2014) conducted a sentence-combining task for 45 native speakers of English in order to investigate if the order of difficulty conforms to the NPAH prediction. The results showed that the intermediate-high level learners preferred to produce SRCs than ORCs. In addition, she also claimed that SRCs were easier than ORCs through the analysis of learner's response accuracy. This shows that the NPAH is applicable to L2 Chinese. Li (2015) conducted a corpus-based study to analyze RC production by speakers of three L1s (i.e. English, Japanese, and Korean) in the HSK corpus (Zhang et al., 2004). The 201 RC sentences they observed show that all three groups of advanced level learners tended to produce more SRCs, and this therefore also supports the NPAH hypothesis.

Based on a review of the aforementioned studies, we find that, even with similar research methods, contradictory results were reported. This leads us to wonder whether the inconsistent results were caused by different L1s or different language proficiency levels. In Sect. 4, we will address this question further.

2.2 Effects of the Animacy of Head Nouns

Aside from the predictive power of the NPAH, analysis based on language processing has provided much insight into the study of RCs in recent years. For example, Traxler et al. (2002) used eye-tracking testing to conclude that an ORC following an animate head noun is more difficult to process, such as 'The mountaineer that the boulder hit', than an inanimate head noun, such as 'The rock which the boy threw.' This shows that the animacy of a head noun is connected to the difficulty of comprehension of an RC. The results of Ozeki and Shirai (2007) also support the effect of animacy. 1005 tokens of Japanese RCs by native speakers of English, Korean, and Chinese were collected from an oral interview corpus. They concluded that English-native and Chinese-native L2 Japanese learners made strong associations between Subject and animate heads and between Direct Object/Oblique and inanimate heads.

There is very limited research on the role of animacy in the L1 Chinese acquisition of RCs. The two most prominent studies on L1 Chinese are Cheng (1995) and Wu (2011). Cheng used elicitation tasks to examine Mandarin-speaking children's (across age groups of three-, four-, and five years old) production of RCs. Her research is based on the semantic hypothesis that an inanimate argument is easier to comprehend. She has shown that, if a head noun is inanimate, participants demonstrate a higher rate of accuracy and that the noun phrase proves easier to understand. And this tendency is more apparent in younger children. Wu (2011) analyzed 331 RCs in a news corpus. Her results show that SRCs contain more animate heads while ORCs contain more inanimate heads. She suggested that the effect of animacy found in the corpus may account for the inconsistent results of previous experimental studies.

Regarding L2 Chinese learning, by observing the HSK corpus, Li (2015) demonstrated that the animacy of nouns in RCs strongly affects the generation of RC types. He also declared that NPAH is secondary to animacy in affecting the production of RC types. However, his research did not take learners' proficiency into account and observed only a limited 201 samples. In our study, we also adopted a corpus approach, but we observed a total of 2,259 samples of RCs representing learners from various L1s and Chinese proficiency levels. Hopefully, this can provide a better profile to settle the dispute among the inconsistent results described above.

2.3 *Effects of Positions in a Matrix Sentence for SRCs and ORCs*

Some cognitive theories posit that center-embedded RCs may interrupt language processing; therefore, they are more difficult to comprehend than those (right- or left-embedded) which occur on the sides of the matrix sentence (Bever, 1970; Kuno, 1974). Mandarin Chinese is considered a left-branching language. An RC is also based on the left-branching structure to always occur before a head noun, which thus causes an embedded structure with an object position such as (2) and (3), but not with a subject position as shown in (4) and (5). In view of this, Chinese RC nominals in the subject position (either SS or SO) should be easier to process than object-position RC nominals (either OS or OO) as shown in the examples below. In addition, Sheldon (1974) also proposed that RCs with the same position and type are easier to comprehend than different structures; that is, SS and OO are easier to comprehend than SO and OS.

(2) *ta bu shi* [*na ge mai shu de ren*] $_{OS}$.
he not be that-Cl buy book DE person
'He is not [the one who bought that book].'

(3) *ta xihuan* [*Zhangsan mai de shu*] $_{OO}$.
he likes Zhangsan buy DE book
'He likes [the books which Zhangsan bought].'

(4) [*mai shu de nage ren*] SS *bu shi wo tongxue.*
buy book DE that-Cl person not be my classmate
'[The one who bought that book] is not my classmate.'

(5) [*Zhangsan mai de shu*] SO *bu jian le.*'
Zhangsan buy DE book not see ASP
'[The books Zhangsan bought] are lost.'

Dai (2010) aimed to understand how the position and type of RCs impact L2 Chinese acquisition. His study concluded that the position factor has no significant effect on producing SRCs or ORCs. Li's observations (2015) showed that the embedded structures of OS and OO are produced more often for English-native learners. Korean-native learners showed no preference, and Japanese-native learners showed the opposite tendency: SO and SS structures were produced more. However, there were no statistically significant differences among the three learner groups' RC production. It seems that current research shows that the position of RCs has no strong preference effect on the selection of RC types. What is curious is why learners do not avoid using more complicated embedding structures. We will clarify this with our statistical results in Sect. 4.

To summarize, most of the previous research on L2 Chinese RC acquisition has been based on experimental methods, either from the viewpoint of universal grammar or language processing. There is still controversy over the results of such research. Our motive is to discover the difficulty of SRCs and ORCs in order to apply the findings within pedagogical grammar. Therefore, we observed and analyzed written texts spontaneously produced by L2 Chinese learners to examine the distribution, position, and animacy effect of RCs.

3 Methodology

3.1 Research Scope

We would like to clarify some terminology and basic syntactic patterns of Chinese RCs before any further discussion. The basic formation of RC nominals in Mandarin Chinese is not different from common noun phrases, except that the modifier must be either a verb phrase or a clause. The modifier precedes a head noun and ends with the relative particle *de,* which connects with the head noun together to form an RC noun phrase. For instance, *na ben shu* 'that book' is the head noun of the RC nominal *wo xihuan de na ben shu* 'that book that I like' and *wo xihuan de* 'that I like' is the modifier. In this research, 'RC' may be used to refer to either the relative clause itself or sometimes the relative clause NP; a distinction between these two referred clauses will not be made if the ambiguity can be resolved by context or differentiation. A modifier of an RC must be understood as either a verb phrase or a clause, but within this intransitive stative verbs and adjectives are excluded. Therefore, NPs such as

congming de nühai 'a smart girl' or *hen hao de keben* 'a good textbook' are not in the category of RC. Furthermore, this study limits the scope of head nouns to only the subjects or objects of verbs of RCs, i.e. the top two roles within NPAH. For instance, even when modifiers are in a verb phrase, the following examples in which head nouns are not in the subject or object roles are excluded: (1) there is an appositional relationship between the modifier and the head noun, such as *women qu Ouzhou lüxing de jihua* 'our plan to travel to Europe'; this is because *jihua* 'plan' is not an argument of this clause; (2) the head noun is part of a clausal subject, as in *lihunlü hen gao de guojia* 'a country with a high divorce rate'; (3) any instance where the head noun is omitted, for example, *wo xihuan ni mai de (shu)* 'I like what you bought (the book)'. To put it simply, only the head noun of an NP is the subject or object argument of an active verb.

3.2 Research Method

Previous research on RC comprehension or generation for the most part has been based on individual experiments in cognitive psychology, such as online sentence generations, grouping linguistic elements together to form a grammatical RC (Wu & Sheng, 2014), or asking learners to combine two sentences into one with an RC construction (Xu, 2014). These methods use designed test questions to accomplish specific research objectives, and results may be used to test a research hypothesis. However, collected samples are often limited because the number of target subjects is constrained by budget and time. Other than experimental design, another solution is to analyze a much larger quantity of authentic language materials provided by a learner corpus. Learners' language use over different proficiency levels can also be regarded as longitudinal profiles. Hence, corpus-based or corpus-driven studies have provided a new avenue for research (Granger, 1998; Douglas, 2001; Ellis & Barkhuizen, 2005: 48; Myles, 2005).

In order to analyze a large quantity of authentic language used by learners, our research also adopts a corpus-based approach. The corpus we used is the TOCFL Learners Corpus (Chang, 2013).[1] This corpus consists of essays written by non-native Chinese-speaking participants who have taken the TOCFL from 2006 to 2012. It contains 1.6 million words from learners of 42 different language backgrounds, including 4,709 essays on 80 topics written by learners from different proficiency levels. The corpus differs from the HSK corpus used by Li (2015) in Mainland China. The TOCFL is an online test that allows participants to directly type their essays into a computer, and the data in the corpus comes from the beginning, intermediate, and high proficiency level learners (CEFR A2-C1). However, the HSK corpus collects the learners' hand-written compositions and only includes essays from advanced proficiency learners (CEFR B2).

[1] Please visit the website http://tocfl.itc.ntnu.edu.tw/(account: tocfl; pwd: demo123).

Therefore, in this research we have the advantage of using a much larger quantity of data from different native language backgrounds and varying proficiency levels in order to investigate if these groups demonstrate any clear differences when producing RCs in Chinese.

3.3 Corpus Data

Linguistic typologist Joseph Greenberg (1963) has noticed that Mandarin Chinese is different from the 30 other VO word order languages. In his analysis, RCs in the other VO languages are formed by placing the head noun to the left of the modifier (i.e. a right-branching structure); however, Chinese follows a VO word order where the head noun is placed on the right (i.e. a left-branching structure). This unique structure is distinct from other languages. Therefore, this study has observed learners whose native languages have different typologies as classified below (Chen, 2007:236). In order to ensure the generalizability of our analysis and to meet our research goals, this study selects two languages from each type, including Japanese and Korean (type 2), Indonesian and Vietnamese (type 3), and English and Spanish (type 4).

Type 1. Left-branching languages with VO word order: Chinese
Type 2. Left-branching languages with OV word order: Japanese and Korean
Type 3. Right-branching languages with VO word order: Thai, Vietnamese, and Indonesian
Type 4. Right-branching (head nouns + RCs) and left-branching (adjectival modifiers + head nouns) with VO word order: English, German, French, Spanish, and Italian.

The following example uses the NP *xuesheng mai de* (*na ben*) *shu* 'The book that the student bought' to exemplify the structure of RCs in each of the six languages. English, Indonesian, Vietnamese, and Spanish all place the head noun on the left, while Japanese, Korean, and Mandarin Chinese place the head noun on the right.

Chinese:	*xuesheng mai de (na ben) shu*	(the head on the right)
	student bought DE (that CL) book	
Japanese:	*gakusei-ga ___ katta hon*	(the head on the right)
	student-NOM bought book	
Korean:	학생이 산 책	(the head on the right)
	student-NOM bought book	
Indonesian:	buku yang siswa beli	(the head on the left)
	book which student buy	
Vietnamese:	cuốn sách học sinh mua	(the head on the left)
	CL book student buy	
English:	the book which the student bought	(the head on the left)

Table 1 Number and distribution of observed compositions for six L1s

	Japanese	English	Korean	Vietnamese	Indonesian	Spanish	Total
B1	530	344	245	152	163	90	1,524
B2	260	122	130	96	112	15	735
Total	790	466	375	248	275	105	2,259

Spanish: El libro que el estudiante compra (the head on the left)
the book which/that the student bought

In order to ascertain whether language proficiency affects learners' RC expressions, this study investigates two proficiency levels in the TOCFL corpus: B1 (CEFR B1 corresponds to the Intermediate-high level in the ACTFL scale) and B2 (advanced level in the ACTFL scale). The data from the B2 levels can be compared with results from previous studies using the HSK corpus of advanced learners (Li, 2015). Table 1 shows that the total number of observed compositions is 2,259 (1,524 for the B1 level and 735 for the B2 level). We can see that the corpus does not provide a balanced distribution of native speakers from each language background in Table 1; this is because there is not a balanced distribution among test participants in the first place. This is especially true of Spanish-speaking B2 learners who account for only 15 compositions. Therefore, this study provides a quantitative analysis of Spanish speakers as a reference rather than an observation of statistical significance.

3.4 RC Markup Principles for the Corpus

Once selected, corpus materials must be manually reviewed to mark the information of each RC. If an RC is applicable to this investigation, it is copied into a separate Excel spreadsheet. Each RC is then tagged with three pieces of information for analysis: (a) type of RC (ORC 'O' or SRC 'S'), (b) position of the RC nominal in the matrix sentence (subject or object position), and c) animacy of the RC head noun (animate '+' or inanimate '−'), as shown in Table 2.

In marking the RCs in the corpus, since the authentic materials are from language learners' interlanguages, more detailed criteria must be defined to judge partially incorrect samples, as shown in (6)–(12). Despite typos (*xi* is omitted in (6)) or incorrect verb usage in (7), the structure of the RC is still apparent in sentences (6) and (7). Since these errors do not jeopardize the judgment of the RCs, they are still marked as RCs in the statistical analysis. The errors for sentences (8)–(10) are respectively caused by lacking the auxiliary verb (*yao)* and the wrong word order position of the adverb (*zui*) as well as the determiner (*you xie*). Since these errors do not affect comprehension, they also count as RCs. However, sentences (11) and (12) lack the main verb of the RC. Though these sentences may still be comprehensible within

Table 2 Markups for sample RCs

Entry no.	RC nominal	Type	Position	Animacy
1	*mei tian yüdao de shiqing* every day encounter DE event 'the things encountered every day'	O	S	–
2	*ding cai de pengyou* order food DE friend 'the friend who ordered food'	S	S	+
3	*bu tai xiguan shuo Yingwen de xuesheng* not very used to speak English DE student 'the students who are not very used to speaking English'	S	O	+
4	*wo ti ni anpai de liang ge xuanze* I for you make DE two CL choice 'the two choices I made for you'	O	O	–

context, they lack a very important element—the verb. To avoid controversy, samples like number (11) and (12) have been excluded.

(6) *tamen xuyao de dong […]* (should be *tamen xuyao de* ***dongxi*** 'those things which they need')

(7) *tamen xiang yao kai de shengyi* (should be *ta men xiang yao* ***zuo*** *de shengyi* 'the business they would like to do')

(8) *wo zuotian xie de ren* (should be *wo zuotian* ***yao*** *xie de ren* 'those who I thanked yesterday')

(9) *wo bixu zui ganxie de ren* (should be *wo* ***zui bixu*** *ganxie de ren* 'those who I must thank the most.')

(10) *chi de you xie dongxi* (should be ***you xie chi de*** *dongxi* 'There are some things to eat.')

(11) *hen duo cong bu yi yang de guojia de xuesheng* (should be *hen duo cong bu yi yang de guojia* ***lai*** *de xuesheng* 'students from many different countries')

(12) *hen duo jianshu de ren* (should be *hen duo* ***hui*** *jianshu de ren* 'many people who know how to fence').

4 Statistical Results and Discussion

In this study, we observed a total of 2,055 RCs, including 1,362 RCs at the B1 level and 693 RCs at the B2 level. This can be compared with Li's (2015) corpus data, in which he investigated only 201 total RCs among English, Korean, and Japanese native speakers, a sample size significantly smaller than ours. Table 3 shows the total number of RCs produced by each of the six learner groups. Japanese learners produced the largest sample size of 563 RCs, but the largest number of occurrences does not indicate the most frequent use due to uneven distribution of compositions across different language groups. The Japanese essays account for the largest portion of the TOCFL corpus, totaling 187 thousand characters from B1 learners and 128

thousand characters from B2 learners. In fact, the highest frequency of RCs is found among native Korean B1 learners as shown in Table 3.

4.1 Difficulty of ORCs or SRCs

The majority of second language acquisition studies support the NPAH accessibility hypothesis that subject-RCs are easier to acquire than object-RCs. In this study, however, we analyzed more than 2,000 RCs and found that more ORCs were produced than SRCs with a statistical significance value of $p < 0.001$. Tables 4 and 5 provide detailed statistics where the number in parentheses indicates the number of tokens. Table 4 shows that, regardless of the mixed language background, all B1 learners consistently produced more ORCs than SRCs with a statistical significance value of $p < 0.01$. However, at the B2 level (see Table 5), there are some variations of the production advantage across different language backgrounds. English- and Korean-speaking B2 learners produced significantly more SRCs than ORCs with a statistical significance value of $p < 0.05$, averaging about 60%. Japanese-speaking B2 learners did not show a significant difference between the uses of SRCs and ORCs with $X^2 = 0.961$ and $p = 0.327$. As for B2 Indonesian-, Vietnamese-, and Spanish-speaking learners, the ORC still maintained an advantage with a statistical significance value of $p < 0.05$; however, Spanish speakers were excluded due to the small sample size. We also observe that there is an overall increase in the use of SRCs as the learners' language proficiency increases.

This is contrary to the corpus-based findings of Li (2015). While that study showed an advantage of SRCs among English, Japanese, and Korean native speakers, it showed no significant difference in the generation of the two types of RCs. While our data from the TOCFL B2 corpus is similar in quality and proficiency level to that of the HSK, our analysis shows similar results among English-speaking learners but opposite results among Japanese- and Korean-speaking learners. In addition, our B2 Vietnamese- and Indonesian-speaking learners produced an average of 60% more ORCs. The result is in contrast to native English learners' preference, despite the fact that these three languages all have head initial NP structures. Therefore, our investigation of the RCs produced by learners of different native languages and proficiency levels does not support the argument that SRCs are easier than ORCs in Mandarin Chinese. On the contrary, low proficiency level learners consistently produced more ORCs. Coincidentally, the same result is found in Chinese L1 acquisition research, which has indicated that the younger the child, the more likely they are to produce an ORC (Chen & Shirai, 2014; Cheng, 1995; Lee, 1992).

As a result, we hypothesize that ORCs are easier to learn in Chinese because more ORCs are produced for lower proficiency learners regardless of their native language types. The dominant factor for this result may be that the word order of RC nominals is the same as that of Chinese simple sentences. After reaching higher language proficiency, no matter what the language background is, learners gradually achieve more native-like expressions. Past Chinese L1 corpus-based studies all show

Table 3 Statistics on RCs

	Japanese		Korean		Vietnamese		Indonesian		English		Spanish		Total RCs
	Char	RC	Char	RC	Char	RC	Char	RC	Char	RC	Char	RC	
B1	187,650	350	92,650	299	57,879	170	60,660	236	131,443	239	33,312	68	1362
B2	128,697	213	67,795	146	54,876	97	34,093	108	66,902	107	6545	22	693

Table 4 The distribution of subject-RCs and object-RCs produced by B1 learners

B1	Japanese	Korean	Vietnamese	Indonesian	English	Spanish	Average
Subject-RC	24% (83)	21% (63)	24% (41)	22% (52)	23% (55)	31% (21)	23% (315)
Object-RC	76% (267)	79% (236)	76% (129)	78% (183)	77% (184)	69% (47)	77% (1,046)

Table 5 The distribution of subject-RCs and object-RCs produced by B2 learners

B2	Japanese	Korean	Vietnamese	Indonesian	English	Spanish	Average
Subject-RC	47% (95)	59% (83)	35% (33)	25% (25)	60% (62)	5% (1)	45% (299)
Object-RC	53% (109)	41% (58)	65% (62)	75% (77)	40% (41)	95% (21)	55% (368)

that the tokens of SRCs occur more often than ORCs (Hsian & Gibson, 2003; Pu, 2007; Tang, 2007), regardless of the genres of the corpus data. This might explain why B2 learners show an increased use of SRCs.

4.2 Effects of Positions of RCs in a Matrix Sentence

Our review of various theories in cognitive linguistics (e.g. Bever, 1970; Kuno, 1974) posits that the structure of the center-embedded RC may reduce processing speeds and make it more difficult to comprehend than an RC placed on the sides of the matrix sentence. This supposes that for languages with NP head final structure like Chinese, SS or SO structures should be easier to process than OO or OS structures, since OO and OS cause embedded structures while SS and SO do not. Does this hypothesis imply that RCs should occur more in a subject position than in an object position? Tables 6 and 7 show that the embedded structures of OO and OS are actually produced more than the non-embedded structures of SS and SO with a statistical significance value of $p < 0.001$. This goes against some theories in cognitive linguistics (e.g. Bever, 1970; Kuno, 1974; Sheldon, 1974). Our data shows that ease of comprehension seems not to equate to ease of production in language processing.

Li (2015) found that English-speaking Chinese L2 learners generated more embedded structures, Japanese-speaking L2 learners generated fewer embedded structures, and Korean-speaking L2 learners use both positions equally. Therefore, he claims that the position of RCs seems not to affect the preference for RC-type generation. However, he also stated that the result did not reach statistical significance due to the limited observation samples. He studied only 201 RC samples taken from learners who are equivalent to the B2 level of the TOCFL scale. On the contrary, we studied 440 samples from these three languages (i.e. Japanese, Korean, and English) and more from other languages (see Table 7). The results show a preference for the

Table 6 RC position and type distribution of B1 learners

B1	Japanese	Korean	Vietnamese	Indonesian	English	Spanish	Total
SS	10% (35)	7.8% (23)	4.7% (8)	7.7% (17)	4.2% (10)	16.2% (11)	44% (601)
SO	38% (133)	39.8% (117)	38.2% (65)	32.8% (77)	34.7% (83)	32.4% (22)	
OS	13.7% (48)	12.6% (37)	19.4% (33)	14.9% (35)	18.9% (45)	14.7% (10)	56% (755)
OO	38.3% (134)	39.8% (117)	37.6% (64)	45.1% (106)	42.2% (101)	36.8% (25)	
Sum	350	294	170	235	239	68	1,356

Table 7 RC position and type distribution of B2 learners

B2	Japanese	Korean	Vietnamese	Indonesian	English	Spanish	Total
SS	16.7% (34)	33.8% (45)	13.7% (13)	17% (17)	16.5% (17)	0% (0)	41% (270)
SO	17.6% (36)	23.3% (31)	23.1% (22)	36% (36)	12.6% (13)	27.3% (6)	
OS	29.9% (61)	24.1% (32)	21.1% (20)	8% (8)	43.7% (45)	4.5% (1)	59% (387)
OO	35.8% (73)	18.8% (25)	42.1% (40)	39% (39)	27.2% (28)	68.2% (15)	
Tokens	204	133	95	100	103	22	657

generation of embedded structures (OO, OS) over non-embedded structures (SS, SO).

So, what is the factor that causes more production of the more difficult embedded structures? Li (2015) provided the following explanations. Since L2 learners lack sufficient language proficiency, they are more inclined to generate simple RCs (p.37) and their processing of short RC NPs resembles that of idiom chunks, which does not cause difficulties for sentence generation. However, while such an explanation might satisfy his claim that 'the position of RCs seems not to affect the preference of RC type generation', it cannot explain why ORCs were generated more regardless of their position. Here, we reassert our previous hypothesis that word order is the dominant factor. Since the word order of ORCs is the same as the word order of basic Chinese sentences, i.e. SVO, it results in learners preferring to generate ORCs because it does not require additional processing effort. Tables 6 and 7 show that ORCs are generated significantly more than SRCs in both the subject position and the object position (i.e. SO and OO being generated more than OS and SS, respectively), no matter whether there is structure embedding or not.

Overall, the patterns of OO and SO have the advantage. The average distribution of OO is 39% and SO is 30%. That means the occurrence of the ORC type is not affected by the position in the matrix sentence. No matter where the ORCs are positioned,

their tokens occur more than SRCs. However, for the B2-level Korean- and English-speaking learners, SRC has an obvious advantage. OS is 43.7% for Korean speakers while SS is 33.8% for English speakers, as shown in Table 7. At present, we do not have a good explanation for this part of the data. The only possible assumption is that the B2 English speakers and Korean speakers produce more SRCs than ORCs as indicated in Table 5; therefore, they show this special tendency.

4.3 Animacy of Head Nouns

Previous research shows that 1) the animacy of a head noun is related to the comprehension of an RC and 2) SRCs tend to modify animate noun phrases while ORCs tend to modify inanimate noun phrases. The animacy effect may affect the distribution of the types of RCs. We can see this tendency clearly in Table 8 where the data shows that SRCs prefer animate heads (overall average is 67%), whereas the ORCs prefer inanimate heads (overall average is 85%). It is also consistent with research done by Ozeki and Shirai (2007). Our data also shows that the association between ORCs and inanimate heads is stronger than between SRCs and animate heads. Such a phenomenon may explain why the lower proficiency language learners produce more ORC types than SRC types. In addition to the factor of word order, the processing of inanimate nouns is easier than that of animate nouns (Cheng, 1995).

Furthermore, the animacy effect on SRCs becomes stronger as language proficiency moves from B1 to B2, regardless of the learners' native language. The overall average of SRCs with animate nouns for the B1 level and the B2 level is 58% and 77%, respectively, with a statistical significance value of $p < 0.005$. Though the overall average of ORCs with inanimate nouns from B1 (83%) to B2 (92%) shows the same tendency, there is no significance ($p = 0.0934$).

For another statistical perspective, Table 9 exhibits the analysis of the relationship between SRCs/ORCs and the animacy of the head noun. In the B1 level, English-native Chinese learners used a total of 85 animate head nouns, with 53 modified by ORCs and only 32 modified by SRCs. The tendency of Indonesian-native learners is the same as English learners, i.e. ORCs have the advantage. Korean-native learners used both structures (object- and subject-RCs) equally and do not exhibit a clear preference for animate head nouns. Japanese-native learners used 103 animate head nouns, with 54 occurring with SRCs. While this number is slightly higher than ORCs, there is only a difference of five RCs. The distribution of Vietnamese- and Spanish-speaking L2 learners is higher among SRCs. It is true that totally animate nouns occur more often with SRCs, but the different distribution is not statistically significant ($p = 0.8357$).

However, for the B2 level, there was an overall average of 89% of animate nouns occurring in the SRCs. Based on the discrepancy between the B1 and B2 levels, we may conclude that once high language proficiency has been reached, the effect of animacy takes over as the dominant factor in RC type generation. For lower proficiency learners, the factor of word order, instead of animacy, is dominant in

Table 8 The distribution of animacy of head nouns modified by RCs

Type	Animacy	Japanese		Korean		Vietnamese		Indonesian		English		Spanish	
SRC	Animate	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
		58% (32)	84% (52)	65% (54)	81% (77)	53% (33)	71% (59)	51% (31)	71% (20)	59% (24)	70% (23)	67% (14)	100% (1)
	Inanimate	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
		42% (23)	16% (10)	35% (29)	19% (18)	47% (29)	29% (24)	49% (30)	29% (8)	41% (17)	30% (10)	33% (7)	0% (0)
ORC	Animate	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
		29% (53)	2% (1)	18% (49)	9% (10)	14% (34)	7% (4)	18% (36)	9% (7)	5% (7)	6% (4)	10% (5)	14% (3)
	Inanimate	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
		71% (131)	98% (40)	82% (218)	91% (99)	86% (202)	93% (54)	82% (164)	91% (71)	95% (122)	94% (58)	90% (43)	86% (18)

Table 9 The distribution of animate and inanimate head nouns of RCs with different types

	B1 animate nouns		B2 animate nouns	
	SRC	ORC	SRC	ORC
English	38% (32)	62% (53)	98% (52)	2% (1)
Japanese	52% (54)	48% (49)	88% (77)	12% (10)
Korean	50% (33)	50% (34)	94% (59)	6% (4)
Indonesian	46% (31)	54% (36)	74% (20)	26% (7)
Vietnamese	77% (24)	23% (7)	85% (23)	15% (4)
Spanish	74% (14)	26% (5)	25% (1)	75% (3)
Average	51% (188)	49% (184)	89% (232)	11% (29)

producing the type of RCs. Findings yielded in Tables 8 and 9 thus lead to the conclusions that (1) there is a strong association between ORCs and inanimate heads; and (2) the lack of significant association between SRCs and animate nouns at the B1 level indicates that ORC type is easier than SRC type for lower proficiency Chinese learners.

5 Conclusion and Limitations of This Study

This study analyzed the L2 Chinese learners' production of RCs among six different native languages classified into three language typologies and discussed the corpus results from multiple perspectives, including (1) the types of RCs, (2) the position of an RC in the matrix clause, (3) animacy of the RC modifying head noun, (4) L2 learners' native languages, and (5) different proficiency levels of learners. Specifically, we examined how those mingling features affect the production of different types of RCs.

We have found that B1 (intermediate-high level) learners produce significantly more ORCs than B2 (advanced level) learners. This trend consistently occurs among all language backgrounds, thus indicating that lower proficiency learners produce more ORCs than SRCs. This phenomenon also occurs in L1 Mandarin Chinese acquisition. Owing to this effect (i.e. more ORCs), the learners produced more OO structure RCs, which is in opposition to findings of previous research on language processing theories, which imply that object position RCs should be produced less due to their embedded structure. Furthermore, the animacy effect for SRCs at the B1 level apparently does not occur either. This leads us to conclude the following hypothesis: at the early stage of L2 Chinese learning, word order is the dominant factor for language processing no matter what the learner's native language is. Since SVO is the conventional word order in Mandarin Chinese and noun phrases modified by ORC have the same SVO structure, this results in the fact that ORCs are easier than SRCs. This also explains why there is no obvious effect on the position feature

causing embedded structure and preference for animate head nouns in SRCs among B1 learners. As learners' proficiency gradually increases, learners produce more SRCs and their interlanguage will approximate the target language.

While a corpus-based study may reflect a learner's natural language production, there are also some limits to this research. For example, the data for Spanish-native learners in this study is just for readers' reference since the corpus does not include a sufficient sample size to produce reliable statistics. Nevertheless, through the use of the spontaneously produced data of learners of different native languages and proficiency levels, the language development of L2 learners can be clearly observed and analyzed. These corpus-based results may be combined with the results from psychological or cognitive linguistic experiments to represent interlanguage development from more comprehensive perspectives.

Acknowledgements The author would like to thank the anonymous reviewers for their valuable comments and the support of the Ministry of Science and Technology, under the grant MOST 108-2410-H-002-117.

References

Bever, T. G. (1970). The cognitive basis for linguistic structures. In: Hayes, J. R. (Ed.), *Cognition and the development of language* (pp. 279–352). New York: Wiley.

Chang, L.-P. (2013). The construction and implication of the TOCFL corpus. In: B. Zhang & X. Cui (Eds.), *The proceedings of the 2nd International Conference of Interlanguage Corpora* (pp. 141–152). Beijing: Beijing Language and Culture University Press. [張莉萍, 2013, 〈TOCFL作文語料庫的建置與應用〉, 崔希亮、張寶林 主編,《第二屆漢語中介語語料庫建設與應用國際學術討論會論文選集》141–152。北京: 北京語言大學出版社。].

Chen, J., & Shirai, Y. (2014). The acquisition of relative clauses in spontaneous child speech in Mandarin Chinese. *Journal of Child Language*. CJO 2014 https://doi.org/10.1017/S0305000914000051.

Chen, F. J. (2007). *Contrastive analysis and its applications in language pedagogy*. Taipei: Crane Publishing. [陳俊光, 2007,《對比分析與教學應用》。台北: 文鶴出版社。].

Cheng, Y.-Y. (1995). The acquisition of relative clauses in Chinese. MA thesis, National Taiwan Normal University, Taipei.

Dai, Y. (2010). An investigation of the relative clause acquisition by learners of Chinese as a second language. *Journal of Ocean University of China (Social Sciences Edition), 6*, 85–91. [戴運財, 2010, 〈漢語作為第二語言的關係從句習得難度調查〉。《中國海洋大學學報》(社會科學版) 6:85–91。].

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition, 13*, 431–469.

Douglas, D. (2001). Performance consistency in second language acquisition and language testing research: A conceptual gap. *Second Language Research, 17*(4), 442–456.

Eckman, F. R., Bell, L. H., & Nelson, D. (1988). On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied Linguistics, 9*, 1–20.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Granger, S. (Ed.). (1998). *Learner English on Computer*. New York: Longman.

Gass, S. (1979). Language transfer and universal grammatical relations. *Language Learning, 29*, 327–344.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of human language* (pp. 73–113). Cambridge, Mass: MIT Press.

Hasegawa, T. (2005). Relative clause production by JSL children. In: M. Minami, H. Kobayashi, M. Nakayama, & H. Sirai (Eds.), *Studies in language sciences 4: Papers from the fourth annual conference of the Japanese society for language sciences* (pp. 189–204). Tokyo: Kurosio.

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*, 3–27.

Izumi, S. (2003). Processing difficulty in comprehension and production of relative clauses by learners of English as a second language. *Language Learning, 53*, 285–323.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry, 8*(1), 63–99.

Kuno, S. (1974). The position of relative clauses and conjunctions. *Linguistic Inquiry, 5*, 117–136.

Lee, T. H.-T. (1992). The inadequacy of processing heuristics–evidence from relative clause acquisition in Mandarin clause. In: Lee, T. (Ed.) *Research on Chinese linguistics in Hong Kong* (pp. 47–85). Hong Kong: The Linguistic Society of Hong Kong.

Li, J. (2015). Research on Chinese relative clause generation: A language type perspective. *Modern Foreign Language Research, 2*, 34–39. [李金滿, 2015, 〈漢語二語關係從句產出研究–類型學視角〉。《當代外語研究》2:34–39。].

Myles, F. (2005). International corpora and second language acquisition research. *Second Language Research, 21*(4), 373–391.

O'Grady, W., Lee, M., & Choo, M. (2003). A subject-object asymmetry in the acquisition of relative clauses in Korean as a second language. *Studies in Second Language Acquisition, 25*(3), 433–448.

Ozeki, H., & Shirai, Y. (2007). Does the noun phrase accessibility hierarchy predict the difficulty order in the acquisition of Japanese relative clauses? *Studies in Second Language Acquisition, 29*(2), 169–196.

Packard, J. L. (2008). Relative clause processing in L2 speakers of Mandarin and English. *Journal of the Chinese Language Teachers Association, 43*(2), 107–146.

Pu, M.-M. (2007). The distribution of relative clauses in Chinese discourse. *Discourse Processes, 43*(1), 25–53.

Sakamoto, T., & Kubota, S. (2000). Nihongo no kankeisetu no syuutoku ni tuite [On the acquisition of Japanese relative clauses]. *Nanzan-Daigaku Kyoiku Sentaa Kiyoo [The Bulletin of the Center for International Education, Nanzan University], 1*, 114–126.

Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior, 13*(3), 272–281.

Tang, Z. (2007). Position of relative clause: An analysis based on corpus and genres. *Modern Lingustics, 2*, 139–150. [唐正大, 2007, 關係化對象與關係從句的位置—基於真實語料和類型分析。《當代語言學》2:139–150。].

Tarallo, F., & Myhill, J. (1983). Interference and natural language in second language acquisition. *Language Learning, 33*, 55–76.

Traxler, M., Morris, R., & Seely, R. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language, 47*, 69–90.

Wu, F. (2011). Experience effect or memory effect?-Evidence from new corpus for Animacy effect of relative clause generation. *Language Sciences*, *53*, 396–408. [吳芙芸, 2011, 〈基於經驗還是基於工作記憶?–來自漢語新聞語料庫中關係從句生命度格局的證據〉。《語言科學》53:396–408。].

Wu, F., & Sheng, Y. (2014). Pre-RC determiner phrase bias and production preference for object relatives: Perspectives from L2-Chinese learners. *Foreign Language Teaching and Research*, *3*, 14–24. [吳芙芸、盛亞南, 2014,〈指量詞的前置優勢及賓語關係從句的產出優勢:漢語二語學習者視角〉。《外語教學與研究》(外國語文雙月刊) 3:14–24。].

Xu, Y. (2014). Evidence of the accessibility hierarchy in relative clauses in Chinese as a second language. *Language & Linguistics, 15*(3), 435–464.

Zhang, B., Cui, X., & Ren, J. (2004). The construction concept of HSK dynamic composition corpus. In: *The Proceedings of Symposium of the Third National Conference on Language Application*

(pp. 544–554). [張寶林、崔希亮、任傑, 2004, ⟨關於“HSK動態作文語料庫”的建設構想⟩,《第三屆全國語言文字應用學術研討會論文集》544–554。].