Foyer
The Journal of Humanities, Social Sciences, and Education

A Corpus-Based Study of Chinese EFL Learners' Use of Relative Clauses in English Writing

Yifan Feng and Suthathip Thirakunkovit
Faculty of Liberal Arts, Mahidol University

Received: June 30, 2025
Revised: December 26, 2025
Accepted: January 16, 2026

Abstract

This study investigates the use of English Relative Clauses (RCs) in the writing of Chinese EFL learners across four proficiency levels (A2, B1-1, B1-2, B2+), focusing on frequency, accuracy, and error types. Based on a corpus of learner essays, results showed an average of 4.647 RCs per 500 words and an overall accuracy rate of 87.16%. Despite this high accuracy, error analysis revealed four main categories: head-noun-related, relativizer-related, gap-related, and sentence-structure-related errors. The most frequent subtype, subject-verb disagreement, often disrupted sentence cohesion. A one-way ANOVA found no significant differences in RC use or error rates across levels, suggesting a developmental plateau. Persistent errors, especially in subject-verb agreement, may stem from L1 transfer, rule overgeneralization, and fossilization. These findings point to the need for targeted instruction to address fossilized patterns. Key recommendations include explicit teaching of RC structures, focused error correction, and practice emphasizing subject-verb agreement and sentence structure. Such interventions may enhance the accuracy and complexity of learners' RC usage.

Keywords

Relative Clauses, ICNALE, Chinese EFL Learners, Error Analysis

Foyer
The Journal of Humanities, Social Sciences, and Education

การศึกษาการใช้ Relative Clauses ในงานเขียนภาษาอังกฤษของผู้เรียนภาษาอังกฤษเป็นภาษาต่างประเทศชาวจีนโดยอาศัยคลังข้อมูลภาษา

อี้ฟาน เฟิง และ สุธาทิพย์ ถิรคุณโกวิท
คณะศิลปศาสตร์ มหาวิทยาลัยมหิดล

บทคัดย่อ

งานวิจัยนี้ศึกษาการใช้อนุประโยคสัมพันธ์ (Relative Clauses – RCs) ในงานเขียนของผู้เรียนภาษาอังกฤษชาวจีนในฐานะภาษาต่างประเทศ (EFL) โดยแบ่งตามระดับความสามารถสี่ระดับ (A2, B1-1, B1-2, B2+) โดยมุ่งเน้นที่ความถี่ ความถูกต้อง และประเภทของข้อผิดพลาดจากคลังข้อมูลเรียงความของผู้เรียน ผลการวิจัยพบว่ามีการใช้อนุประโยคสัมพันธ์เฉลี่ย 4.647 ครั้งต่อ 500 คำ และมีอัตราความถูกต้องโดยรวมอยู่ที่ 87.16% แม้อัตราความถูกต้องจะอยู่ในระดับสูง แต่การวิเคราะห์ข้อผิดพลาดพบว่า มีข้อผิดพลาดหลักอยู่ 4 ประเภท ได้แก่ ข้อผิดพลาดที่เกี่ยวกับคำนามหลัก (head-noun), คำเชื่อมอนุประโยค (relativizer), ช่องว่างในโครงสร้างประโยค (gap), และโครงสร้างของประโยคโดยรวม โดยประเภทที่พบบ่อยที่สุดคือความไม่สอดคล้องระหว่างประธานกับกริยา (subject-verb disagreement) ซึ่งส่งผลกระทบต่อความต่อเนื่องของประโยคอย่างชัดเจน

การวิเคราะห์ความแปรปรวนแบบทางเดียว (one-way ANOVA) พบว่าไม่มีความแตกต่างอย่างมีนัยสำคัญในด้านการใช้ RC หรืออัตราข้อผิดพลาดระหว่างระดับความสามารถทางภาษา แสดงถึงการหยุดชะงักในการพัฒนาของภาษา (developmental plateau) ข้อผิดพลาดที่ยังคงมีอยู่ โดยเฉพาะในเรื่อง subject-verb agreement อาจมีสาเหตุมาจากการถ่ายโอนจากภาษาแม่

(L1 transfer), การสรุปกฎเกินจริง (rule overgeneralization), และการตกผลึกของข้อผิดพลาด (fossilization) ผลการศึกษานี้ชี้ให้เห็นถึงความจำเป็นในการสอนแบบมุ่งเน้นเป้าหมาย เพื่อแก้ไขรูปแบบข้อผิดพลาดที่ฝังแน่น คำแนะนำหลัก ได้แก่ การสอนโครงสร้าง RC ที่ถูกต้องอย่างชัดเจน การแก้ไขข้อผิดพลาดอย่างตรงจุด และการฝึกฝนที่เน้นความถูกต้องของ subject-verb agreement และโครงสร้างประโยค ซึ่งการแทรกแซงเหล่านี้อาจช่วยเพิ่มความถูกต้องและความซับซ้อนในการใช้ RC ของผู้เรียน

คำสำคัญ

Relative Clauses, ICNALE, analysis ผู้เรียนภาษาอังกฤษเป็นภาษาต่างประเทศชาวจีน, การวิเคราะห์ข้อผิดพลาด

Introduction

Relative clauses (RCs) are essential tools for post-modification in English, offering a level of explicitness that other post-modification devices cannot match, and effectively reducing cognitive load during language processing (Biber et al., 2011; Parkinson & Musgrave, 2014; Yamashita, 1994). RCs are widely used by both native English speakers and English as a Foreign Language (EFL) learners, with their frequency of use generally increasing as learners' proficiency improves (Deveci & Nunn, 2018; Fox & Thompson, 2007). While RCs enhance communication, their embedded structure and variety of forms make them more complex than simple sentence constructions, presenting particular challenges for EFL learners (Alotaibi, 2016; Yamashita, 1994).

A substantial body of research has explored RC errors made by EFL learners from various linguistic backgrounds, including Mandarin Chinese speakers (Bao, 2015; Lin & Chuang, 2014). In China, students typically begin learning RCs in Grade 8, yet they often struggle with their structure, as evidenced by the high error rates in writing assessments (Bao, 2015). While RCs facilitate communication, their complexity often leads to significant difficulties in mastering their use.

However, as Chang (2004) pointed out, mastering RC formation requires more than knowing which relative pronoun (or adverb) to use. Bao's (2015) study, for example, focused primarily on errors involving relativizer selection, which, while significant, do not fully capture the breadth of challenges students face when producing RCs in writing. These challenges are likely to vary depending on the learner's proficiency level, suggesting a need for more nuanced error analysis (Carlsen, 2012). Carlsen (2012) further emphasized that language proficiency not only affects grammatical accuracy but also influences vocabulary and syntactic complexity. However, Bao (2015) treated all students as a homogeneous group, overlooking how students at different proficiency levels might exhibit distinct error patterns in authentic language use. This concern is consistent with Myles (2002),

who argues that learners' varying command of the target language influences the kinds of structural errors they make, as both cognitive and social factors shape their written output. In addition, much of the existing research has been conducted in decontextualized tasks, such as isolated grammar drills, sentence-completion items, or RC-specific test questions, where learners respond to prompts rather than generate language spontaneously. Such settings, while useful for controlled comparisons, may not fully reflect the complexity of learners' natural written production (Yang, 2020).

To address these gaps, the present study uses a corpus-based approach to analyze RC errors in the writing of Chinese EFL learners. By drawing on the International Corpus Network of Asian Learners of English (ICNALE), this study aims to identify the most common error types and their underlying causes, offering a more accurate representation of learner language across different proficiency levels (Granger, 2003; McEnery et al., 2006). Unlike prior studies that focused narrowly on relativizer misuse, this research extends the scope to include errors related to RC formation, such as omissions, misplacements, and structural issues. This approach is particularly significant because RCs remain a persistent challenge in Chinese EFL writing classrooms, despite years of instruction, and existing findings from controlled tasks may not adequately inform pedagogy. By analyzing a large learner corpus, this study captures spontaneous written production, thereby offering a more reliable picture of learners' actual difficulties. Moreover, this comprehensive approach sets clear standards for identifying errors and avoids the subjectivity often associated with spoken language assessments (McEnery et al., 2006; Thirakunkovit, 2021). By doing so, the study seeks to provide more precise insights into the nature of RC errors and the factors that influence them in written English.

Research Objectives

There are three objectives in this study:

(1) To investigate a comprehensive description of Chinese EFL learners' use of RCs, with particular attention to the selection and application of relativizers.

(2) To identify and categorize the common error types in RC production and analyze their frequencies.

(3) To investigate the differences in the frequency of RC usage and the rate of errors across four proficiency levels.

Literature Review

Relative Clauses

RCs are subordinate clauses that postmodify noun phrases (NPs) and are typically introduced by relativizers such as "who" or "which" (Andrews, 2007; Leech & Svartvik, 2003). According to Biber et al. (1999), RCs consist of three components: the head noun, the relativizer, and the gap. They categorize RCs into Restrictive Relative Clauses (RRCs), which define or limit the head noun (e.g., "The study that was conducted last year..."), and Non-Restrictive Relative Clauses (NRRCs), which add supplementary information and are marked by commas (e.g., "The conference, which was held in London..."). Sentence-relative clauses (SRCs), introduced by Leech and Svartvik (2003), apply to entire propositions rather than individual nouns. These distinctions are directly relevant to learner difficulties, as misuse can obscure meaning and lead to errors (Bao, 2015; Lin & Chuang, 2014).

RCs are especially valuable in academic writing, where they allow for the efficient integration of complex information (Abdolmanafi & Rahmani, 2012). For example, "The research paper, which was published in a prestigious journal, presents groundbreaking findings" not only conveys essential context but also enhances the credibility of the claim. Each element of the RC contributes to meaning; misuse, such as confusing RRCs and NRRCs,

can obscure meaning or cause misinterpretation. Therefore, mastery of RCs is essential for syntactic precision and communicative clarity, particularly in written academic contexts.

A Comparison of English and Chinese RCs

Chinese also features RCs composed of a head noun, a relativizer ("de"), and a gap, similar to English RCs in structure but different in form and positioning. One of the most striking contrasts lies in word order: Chinese RCs are prenominal, or left-branching, meaning the clause precedes the head noun, whereas English RCs are postnominal (Hsiao & Gibson, 2003; Li et al., 2016). Here is an example:

Wo3 mei2you3 ni3 xiang3yao4 de xiao1xi.

I don't have you want (RC marker) information.

"I don't have the information that you want."

These structural contrasts contribute to typical learner errors. The limited relativizer inventory in Chinese, mainly "de", can lead to confusion in English, where learners must select from a range of relativizers like "who", "which", or "that" (Richards, 1971). Additionally, Chinese verbs do not vary according to subject number or person, unlike English, where subject-verb agreement is required (Li & Thompson, 1989). Here is an example:

Ting2 zai4zhe4li de na4xie1 che1 bu2jian4le.

Parked here (RC marker) those cars are gone.

"Those cars which were parked here are gone."

the verb "ting" remains unchanged regardless of the plural subject "cars". This often leads to subject-verb disagreement in English writing, such as using "were" with plural subjects. These differences, shaped by L1 structures, are central to understanding the recurring error types in Chinese EFL learners' use of RCs.

Theoretical Frameworks

The acquisition of English RCs by Chinese learners can be understood through three key theoretical lenses: error sources, developmental processes, and cognitive mechanisms. Error Analysis (EA), as proposed by Corder (1967), frames learner errors not as failures but as indicators of underlying acquisition processes. Corder's distinction between interlingual (L1 transfer) and intralingual (L2 rule-based) errors explains many of the common issues in RC production. Interlingual errors arise when structural differences between Chinese and English, such as left-branching versus right-branching RCs, lead to word order problems or the misplacement of relativizers. Intralingual errors, on the other hand, reflect learners' misapplication of L2 rules, for instance by overgeneralizing "that" to all contexts or omitting required elements like relativizers.

Beyond sources of errors, developmental perspectives highlight interlanguage fossilization, where incorrect forms become resistant to change despite further input (Richards & Schmidt, 2013). In RCs, this may manifest in the persistent misuse of relativizers or rigid word order, often due to repeated use without corrective feedback. Selinker (1972) also notes that overgeneralization, applying an L2 rule too broadly, can lead to errors. For example, learners may use "that" in non-restrictive clauses, where it is grammatically incorrect. These frameworks, taken together, offer a comprehensive understanding of how structural and cognitive factors interact in RC acquisition. They underscore the importance of targeted instruction to address entrenched patterns and refine learners' grammatical accuracy.

Previous Studies

Research on relative clause (RC) errors among EFL learners generally falls into two categories: grammatical structure and the learnability of RCs. Kusdianty (2016) reported high error rates in the use of English relative pronouns among Indonesian learners, attributing them to the lack of equivalent forms in Indonesian.

Erdogan (2005) found that Turkish learners often confused restrictive and non-restrictive RCs, resulting in errors with pronoun choice and comma placement. Alotaibi (2016) observed that Arabic-speaking students struggled with resumptive pronouns and passive forms, reflecting syntactic contrasts between Arabic and English. Abdolmanafi and Rahmani (2012) noted that Persian learners frequently misused relative markers and retained redundant pronouns, linked to Persian's flexible RC structures. Similarly, Termjai (2021) found that Thai EFL learners often omitted commas and showed frequent errors in subject-verb agreement within RCs, indicating a gap in their understanding of restrictive vs. non-restrictive clauses. These findings collectively underscore how diverse L1 structures shape learners' RC acquisition and error patterns, emphasizing the need for language-specific instructional approaches.

For Mandarin Chinese speakers, the structural divergence between Chinese and English RCs has been consistently shown to cause persistent errors. Lin and Chuang (2014) observed frequent omission of relative pronouns and redundancy due to interference from Chinese word order. Bao (2015) emphasized that the absence of articles and clear relativizer distinctions in Chinese leads to confusion in English RC usage. Ping (2022) found that Chinese high school learners had error rates of 60–70% across various relativizers, suggesting both transfer effects and limited grammatical awareness. These findings underscore the need for pedagogical strategies that not only address specific error patterns but also help learners understand the syntactic differences between their L1 and English.

Despite the valuable insights provided by these studies, several limitations remain. Most of the research, particularly those focusing on Chinese native speakers, has been conducted in decontextualized settings. This means that many studies may not fully capture learners' RC errors in spontaneous language use. Furthermore, many studies do not distinguish learners by proficiency level, making it difficult to ascertain how RC errors evolve as learners progress. To address these gaps, this study aims to investigate Chinese EFL learners' RC errors in a contextualized situation, while also differentiating students by proficiency level. Moreover, the study

will explore the underlying causes of these errors and propose pedagogical strategies to improve RC instruction based on a deeper understanding of learners' specific needs.

Methodology

Corpus and Data Preparation

This study draws on data from the ICNALE Written Essays, a subcorpus of the International Corpus Network of Asian Learners of English (ICNALE), which includes argumentative essays written under controlled conditions by college-level EFL learners across Asia (Ishikawa, 2013, 2018). The ICNALE is notable for its standardization of task conditions (topic, time, length), and for categorizing learners based on CEFR-aligned English proficiency levels: A2, B1-1, B1-2, and B2+.

The ICNALE Written Essays module contains 5,600 essays written on two topics: Topic A, "The importance of part-time jobs for college students" and Topic B, "The ban on smoking in restaurants". Each essay is limited to 200-300 words and completed in 20-40 minutes without access to external references.

For this study, only data from the CHN section (Chinese learners) were selected, comprising 800 essays written by 400 participants, evenly distributed between the two assigned topics. These learners were categorized across four CEFR proficiency levels based on their reported standardized test scores and vocabulary size.

The data were grouped by proficiency level and topic, forming eight subgroups. Minor formatting issues, such as paragraph break inconsistencies, were corrected during the cleaning process, while typographical errors were preserved to maintain the authenticity of learner language. Table 1 summarizes the final dataset used for analysis.

Data analysis

All essays were imported into AntConc, where the Concordance Tool was used to filter sentences containing potential relativizers. Since some relativizers (e.g., “that”) serve multiple grammatical roles, all retrieved sentences were manually reviewed to ensure the presence of genuine RCs. Sentences with zero relativizers were also manually identified. After screening, all valid RC sentences were collected for analysis.

RCs were analyzed from two perspectives: (1) types of RCs used and (2) errors in RC usage. Sentences were categorized by English proficiency level (A2, B1-1, B1-2, B2+) and type of relativizer (“that”, “which”, “who”, “whom”, “whose”, “zero relativizer”, “where”, “when”, “why”).

Errors were identified according to the grammatical rules outlined by Biber et al. (1999) and Leech and Svartvik (2003), and grouped into four main categories based on RC components:

1. Head-noun-related errors, involving omitted head nouns or subject-verb disagreement.

Table 1*: The Detail of the Learners and Essays in the CHN Section*

English proficiency level	Number of EFL learners	Topic	Number of essays	Number of words	
				Each	Total
A2	50	PTJ	50	10,969	21,950
		SMK	50	11,082	
B1-1	232	PTJ	232	52,524	107,018
		SMK	232	52,784	
B1-2	105	PTJ	105	26,755	50679
		SMK	105	24,721	
B2+	13	PTJ	13	3,324	6,369
		SMK	13	3,159	
Total	400	---	800		186,016

2. Relativizer-related errors, including misused or omitted relative pronouns.

3. Gap-related errors, involving missing or incorrectly placed gaps.

4. Sentence-structure-related errors, where RCs are misused entirely.

Detailed examples for each error category are presented in the Findings section.

Only errors within RCs were considered; errors from main clauses or unrelated lexical issues were excluded.

To ensure reliability, 10% of the data were independently coded by a co-rater with 11 years of teaching experience. Cohen's Kappa showed high inter-rater reliability (κ = .837, 95% CI [.810, .990], $p < .001$).

Accuracy rates and error frequencies were calculated for each proficiency group. For comparisons across levels, a one-way ANOVA with post-hoc tests was conducted using SPSS. Statistical significance was set at $p < .05$.

Results

Overview of the Data

After data cleaning, a total of 1,729 RCs were identified across 800 essays by Chinese learners in the ICNALE Written Essays corpus. Among them, 222 instances were classified as incorrect based on the grammatical frameworks by Biber et al. (1999) and Leech and Svartvik (2003).

To account for variations in essay length, all frequencies were normalized to a per-500-word basis. The average RC frequency was 4.647 per 500 words, with an average error rate of 0.597 incorrect RCs per 500 words. Tables 2 and 3 show the distribution of RCs and their errors across the four proficiency levels.

Table 2 shows that the frequency of RCs increases with proficiency: learners at the B2+ level produced the highest RC frequency (6.045 per 500 words), while A2 learners used the fewest (3.964).

Table 3 indicates that A2 learners had the highest error frequency (0.752 per 500 words), while B2+ learners had the lowest (0.550), suggesting a slight decrease in errors with increasing proficiency.

Table 2: ***Numbers and Frequencies of RCs in the CHN Section***

Proficiency Level	Raw Number of RCs	Frequency of RCs (per 500 words)
A2	174	3.964
B1-1	997	4.658
B1-2	481	4.746
B2+	77	6.045
Total	1729	4.647

Table 3: ***Numbers and Frequencies of Errors in the CHN Section***

Proficiency Level	Raw Number of Errors	Frequency of Errors (per 500 words)
A2	33	0.752
B1-1	129	0.603
B1-2	53	0.523
B2+	7	0.550
Total	222	0.597

Research Objective 1

To address the first research objective, the frequency and accuracy of RCs with different relativizers were examined across all essays. Frequencies were normalized per 500 words to ensure comparability.

Table 4 presents the raw numbers, proportions, normalized frequencies, error rates, and accuracy rates of nine English relativizers used by Chinese learners. “Which” (33.95%) and “Who” (32.56%) were the most frequent relativizers, while “Whom” (0.23%) was the least frequent. The overall accuracy rate of RC use was 87.16%.

Table 4: ***Frequencies and Accuracy Rates of Relativizers***

Relativizer	Raw number	Proportion	RCs per 500 words	Errors per 500 words	Accuracy Rate
Which	587	33.95%	1.578	0.255	83.82%
Who	563	32.56%	1.513	0.175	88.45%
Whom	4	0.23%	0.011	0.003	75.00%
Whose	20	1.16%	0.054	0.003	95.00%
That	260	15.04%	0.699	0.121	82.69%
Zero	176	10.18%	0.473	0.000	100.00%
Where	91	5.26%	0.245	0.030	87.91%
When	8	0.46%	0.022	0.000	100.00%
Why	20	1.16%	0.054	0.011	80.00%
Total	1729	100.00%	4.647	0.597	87.16%

These results suggest that Chinese learners demonstrated higher accuracy with certain relativizers such as “Zero”, “When”, and “Whose”, while encountering greater difficulties with “Whom”, “Why”, “That”, and “Which”.

Research Objective 2

Research Objective 2 aims to identify the common types and frequencies of errors that Chinese EFL learners made. A total of 222 errors were identified in the CHN section of the ICNALE Written Essays. These errors were classified into four main categories: head-noun-related errors, relativizer-related errors, gap-related errors, and sentence-structure-related errors. Each category was further divided into ten specific error types. Table 5 presents the raw counts and normalized frequencies per 500 words.

Among these categories, sentence-structure-related errors were the most frequent, occurring at a rate of 0.210 errors per 500 words, while gap-related errors were the least frequent, with a rate of 0.013. In terms of specific error types, subject-verb disagreement was the most common, with a frequency of 0.180. Conversely, repetitive use of relativizers and lack of RC were the least frequent error types, both occurring at a rate of 0.005 per 500 words. All the types are explained and exemplified below with the proficiency level it was found:

(1) Head-noun-related error

1) Lack of head noun: Missing a head noun in the RC.

(S1: A2)[1] …we will have a chance to see what the society is earlier than * who don't take part-time jobs. (A head noun should be added at the place of the * symbol, such as "those".)

Table 5: ***Detailed Error Classification and Frequencies***

Category and Subtype	Raw Number	Frequency (per 500 words)
Head-noun-related error	71	0.188
Lack of head noun	4	0.011
Subject-verb disagreement	67	0.180
Relativizer-related error	68	0.185
Repetitive use of relativizer	2	0.005
Missing relativizer	32	0.086
Wrong relativizer	34	0.091
Gap-related error	5	0.013
No gap in clause	5	0.013
Sentence-structure-related error	78	0.210
Improper use of RC	21	0.056
Lack of RC	1	0.003
Misuse of RRC and NRRC	35	0.094
Wrong sentence structure in clause	21	0.056
Total	222	0.597

(S2: B2+) And when we find out that the rest 1/4 is * who that smokes... (A head noun should be added at the place of the * symbol, such as “people”. The error in “that” will be discussed in another type of error)

2) Subject-verb disagreement: The verb does not agree with the head noun.

(S3: A2) So some special places which is for smoking must be built in the restaurant. (The underlined part should be “are”)

(S4: B1-2) The offensive smell fills the restaurant, which affect people’s appetite. (The underlined part should be “affects”.)

(2) Relativizer-related error

1) Repetitive use of relativizer: Two relativizers are used simultaneously.

(S5: B1-2) I think one of the reasons that why there are so many people being addicted to smoking is people's following the examples of each other. (The underlined part should be deleted.)

(S6: B2+) And when we find out that the rest 1/4 is (people) who that smokes, we have to acknowledge the dangerous situation… (The underlined part should be deleted.)

2) Missing relativizer: A necessary relativizer is absent when it is required in the RC.

(S7: A2) If you can't quit smoking, you should find a place * belongs to you for smoking. (A relativizer should be added at the place of the * symbol, such as "that".)

(S8: B1-1) Many students * have a part-time job get a lot of experience form it and they find a good job later. (A relativizer "who" should be added at the place of the * symbol.)

3) Wrong relativizer: An inappropriate relativizer is used.

(S9: B1-1) They will get other skills what the same old person doesn't know by the day past. (The underlined part should be "that".)

(3) Gap-related error

1) No gap in clause: The RC lacks a necessary gap.

(S10: B1-1) What is more, smoke can cause all kinds of disease that it may make a man die. (The underlined part should be deleted.)

(S11: B1-2) During the process that we try it, a new skill will be owned by us. (The underlined part should be deleted.)

(4) Sentence-structure-related error

1) Improper use of RC: An RC is wrongly used in a context that requires another structure.

(S12: A2) Some people agree with the idea of topic A, which part-time job is very good for college students. (The underlined part should be "that".)

(S13: B1-1) Smoking, a heavy topic, which is always accompanied with sad stories. (The underlined part should be deleted.)

2) Lack of RC: An RC should be used, but it is not used.

(S14: B1-2)[2] On the other hands doing a part-time job can bring some money which will decrease the payment of the parents, it is significant for students who was bore in poverty family. (The underlined part should be "which".)

3) Misuse of RRC and NRRC: Confusion between RRCs and NRRCs

(S15: A2) Not only are more people join the smoking family, but the number of people, who are smoked, is rising rapidly. (The underlined comma should be deleted.)

(S16: B1-1) Smoking in public places can lead to more people begin to smoke especially young people * which is bad for our society. (A comma should be added at the place of the * symbol.)

4) Wrong sentence structure in clause: The structure within the RC is grammatically incorrect.

(S17: A2) So, they can take a part-time job, from which * improve their ability and accumulate experience. (A subject of the clause should be added at the place of the * symbol, such as "they".)

(S18: B1-1) First, it do harm to anyone who * around you. (A verb of the clause should be added at the place of the * symbol, such as "is".)

Research Objective 3

Research Question 3 examines whether there is a difference in the frequency of RC usage and error rates across the four proficiency levels. To address this question, statistical analysis was performed using SPSS, with individual data normalized to a per-500-word scale for each essay.

In order to investigate the relationship between frequency of RC usage and student's proficiency level, a one-way ANOVA was applied. The means and the standard deviations are presented in Table 6.

Table 6: ***Descriptive Statistics for Frequency of RC Usage***

Proficiency Level	M	SD
A2	3.98	2.11
B1-1	4.60	2.54
B1-2	4.75	2.8
B2+	5.88	3.34

The ANOVA was not significant at the 0.05 level, $F(3, 396) = 2.42$, $p = 0.101 > 0.05$, indicating that the frequency of RC usage did not vary significantly across proficiency levels.

The error rate for RC usage was calculated as the ratio of error frequency to the frequency of RC usage. However, 14 students did not employ RCs in their essays, rendering their error rates unidentifiable. Consequently, the valid sample size was N = 386, distributed across the following groups: A2 (N = 49), B1-1 (N = 224), B1-2 (N = 101), and B2+ (N = 12).

A one-way ANOVA was also employed to evaluate the relationship between error rate and student's proficiency level. The means and the standard deviations are presented in Table 7.

Table 7: ***Descriptive Statistics for Error Rate***

Proficiency Level	M	SD
A2	0.19	0.30
B1-1	0.14	0.22
B1-2	0.10	0.18
B2+	0.07	0.10

The ANOVA was not significant at the 0.05 level, $F(3, 382) = 2.42$, $p = 0.065 > 0.05$, suggesting no statistically significant differences in error rates among proficiency levels.

Although the ANOVA results did not reach statistical significance, the descriptive statistics reveal a clear trend: as proficiency level increased, learners tended to employ RCs more frequently and with fewer errors. Specifically, A2 learners showed the lowest frequency of RCs but the highest error rate, whereas B2+ learners produced RCs most frequently and with the fewest errors. This suggests a developmental pattern in RC usage that aligns with general expectations of proficiency growth. However, the small sample size in the B2+ group (N = 12) may have reduced the statistical power of the analysis. Future research with a more balanced distribution of proficiency levels could further examine this tendency.

Conclusion

This study investigated the use and accuracy of English RCs among Chinese EFL learners based on the ICNALE Written Essays corpus. Three research objectives guided the analysis: examining RC usage across relativizers, identifying common RC error types, and exploring proficiency-level differences in RC frequency and accuracy.

The results revealed that learners produced a total of 1,729 RCs across all proficiency levels, with an overall accuracy rate of 87.16%. The relativizers “zero” and “where” showed the highest accuracy, while “whom” and “why” were among the least accurate. A total of 222 RC-related errors were identified and classified into

four main categories: head-noun-related errors, relativizer-related errors, gap-related errors, and sentence-structure-related errors. The most common specific error types were subject-verb disagreement, wrong relativizer, misuse of RRC and NRRC, and missing relativizer.

Statistical analysis revealed no significant differences in RC usage or error rates across proficiency levels, though a downward trend in error rates was observed with increasing proficiency. These findings suggest that while learners gain stability in RC usage over time, accuracy does not always improve proportionally, possibly due to fossilization and avoidance strategies.

Discussion

Interpretation of Findings

The comparatively high accuracy rate in this study, in contrast to previous research using decontextualized tasks (e.g., Bao, 2015; Ping, 2022), highlights the influence of task type on learner performance. Whereas earlier studies often required learners to complete controlled RC-specific items, the ICNALE corpus captured learners' spontaneous production in argumentative essays. This contextualized setting allowed learners to self-select simpler structures, possibly leading to a higher apparent accuracy—a phenomenon related to avoidance strategies (Ellis, 1994).

The ranking of relativizer accuracy also deviates from prior research. For example, while "why" and "when" had high accuracy in some previous studies, this study found "zero" and "where" to be more accurately used. These discrepancies reflect the different linguistic and cognitive demands posed by contextualized versus decontextualized tasks and underscore the importance of authentic writing contexts in studying interlanguage development.

Error analysis confirmed the strong influence of L1 transfer, particularly with respect to subject-verb agreement (which is absent in Chinese), the overuse or misuse of relativizers (due to the singular Chinese marker "de"), and confusion between RRCs and NRRCs (possibly due to the absence of commas in Chinese RCs).

The persistence of these errors across proficiency levels suggests that some may be fossilized, especially in light of the limited improvement observed in more advanced learners.

Implications of the Study

These findings carry practical implications for language instruction. First, teaching should prioritize commonly recurring RC errors, such as subject-verb disagreement and inappropriate relativizer use. Exercises should focus on syntactic elements that contrast sharply with Chinese, such as restrictive/non-restrictive distinctions and proper relativizer selection.

Second, instruction should be contextualized rather than isolated. Grammar exercises embedded in authentic writing tasks are more likely to reflect learners' actual performance and highlight avoidance patterns that might not emerge in decontextualized settings. Teachers should also be aware of fossilized patterns and respond with immediate corrective feedback and targeted intervention.

For learners, the study provides insight into how Chinese linguistic features may hinder English RC acquisition and encourages them to consciously monitor and address fossilized errors in their writing.

For researchers, this study demonstrates the value of corpus-based analysis in examining RC development. The findings reaffirm the role of L1 transfer and avoidance strategies in shaping learner output. They also raise questions about the plateauing of RC accuracy at intermediate proficiency levels and the need to explore fossilization in greater depth.

Limitations and Recommendations for Future Research

Despite its strengths, this study has several limitations. The participant sample was unevenly distributed across proficiency levels, with a heavy concentration at B1-1 and very few learners at B2+, possibly affecting the robustness of comparative analysis.

Furthermore, only written data were analyzed; learners' spoken RC usage remains unexplored.

Another limitation lies in the scope of RC types and relativizers analyzed. Since these were not pre-controlled and the data came from a specific corpus, the findings cannot be generalized to all learner populations. However, this design was intentional: the use of a contextualized task allowed learners to freely employ RCs, thereby reflecting more authentic and less constrained language use. This contextualized perspective, while less generalizable, offers unique insights that decontextualized tasks cannot capture.

Future research should aim to balance proficiency-level samples and include oral production data to gain a more comprehensive view of RC acquisition. In addition, more controllable designs, perhaps combining contextualized and decontextualized tasks, can generate findings that are both universal and reflect natural learner performance. Longitudinal studies could also help track changes in RC usage and accuracy over time, providing deeper insight into learner development and fossilization.

Note

1. S1 means Sentence 1. A2 shows the proficiency level of the student who wrote this sentence. The proficiency levels include A2, B1-1, B1-2, and B2+. The same convention applies throughout the examples from ICNALE.
2. This case is classified as Lack of RC because the learner appeared to intend a relative clause. Although both "revising it to an NRRC" and "dividing it into two independent sentences" would be grammatically acceptable revisions, the analysis adopts the former option for consistency.

References

Abdolmanafi, S. J., & Rahmani, Z. (2012). An investigation of the learnability of relative clauses by EFL learners. *World Journal of English Language, 2*(3), 29–37.

Alotaibi, A. M. (2016). Examining the learnability of English relative clauses: Evidence from Kuwaiti EFL learners. *English Language Teaching, 9*(2), 57–65.

Andrews, A. D. (2007). Relative clauses. In T. Shopen (Ed.), *Language typology and syntactic description. Vol. 2, Complex constructions* (2nd ed). (pp. 206–236). Cambridge University Press.

Bao, X. (2015). Senior high school students' errors on the use of relative words. *English Language Teaching, 8*(3), 137–154.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. Longman.

Carlsen, C. (2012). Proficiency level—a fuzzy variable in computer learner corpora. *Applied Linguistics, 33*(2), 161–183.

Chang, J. (2004). *Examining models of second language knowledge with specific reference to relative clauses: A model – comparison approach* [Doctoral dissertation, Teachers College, Columbia University].

Corder, S. P. (1967). The significance of learner's errors. *IRAL – International Review of Applied Linguistics in Language Teaching, 5*(4), 161–170.

Deveci, T., & Nunn, R. (2018). The use of relative clauses in humanities and social sciences research articles: A case study. *Linguistics and Literature Studies, 6*(1), 17–26.

Ellis, R. (1994). *The study of second language acquisition.* Oxford University Press.

Erdogan, A. G. V. (2005). Use of English relative clauses by Turkish learners: A study of errors. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi, 3*(2), 22–28.

Fox, B. A., & Thompson, S. A. (2007). Relative clauses in English conversation: Relativizers, frequency, and the notion of construction. *Studies in Language, 31*(2), 293–326.

Granger, S. (2003). The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538–546.

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition, 90*(1), 3–27.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world, 1*(1), 91–118.

Ishikawa, S. (2018). The ICNALE edited essays; A dataset for analysis of L2 English learner essays based on a new integrative viewpoint. *English Corpus Studies, 25*, 117–130.

Kusdianty, S. R. (2016). The analysis of students' error in using relative clause. *Indonesian EFL Journal, 2*(1), 60–70.

Leech, G., & Svartvik, J. (2003). *A communicative grammar of English* (3rd ed.). Routledge.

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar.* University of California Press.

Li, Q., Guo, X., Yao, Y., & Müller, N. (2016). Relative clause preference in learners of Chinese as a second language. *Chinese Journal of Applied Linguistics, 39*(2), 199–214.

Lin, C. Y., & Chuang, S. Y. (2014). Taiwanese EFL learners' English relative clauses learning. *International Journal of English Linguistics, 4*(4), 99–106.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus – based language studies: An advanced resource book.* Routledge.

Myles, J. (2002). Second language writing and research: The writing process and error analysis in student texts. *Tesl–ej*, *6*(2), 1–20.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for academic purposes students. *Journal of English for Academic Purposes, 14*, 48–59.

Ping, W. (2022). The analysis of Chinese intermediate students' error in using relative clauses. *Frontiers in Educational Research, 5*(3), 73–78.

Richards, J. (1971). *Error analysis and second language strategies* (Report No. ED048579). ERIC. relativizers https://eric.ed.gov/?id=ED048579

Richards, J. C., & Schmidt, R. W. (2013). *Longman dictionary of language teaching and applied linguistics.* Routledge.

Selinker, L. (1972). Interlanguage. *IRAL: International Review of Applied Linguistics in Language Teaching, 10*(1–4), 209–232.

Termjai M. (2021). Error analysis of English relative clauses in argumentative essays written by Thai EFL learners. *Journal of Humanities, Naresuan University, 18*(3), 87–105.

Thirakunkovit S. (2021). *An introduction to corpus linguistics: The use of corpora for English language teaching and learning.* Printery.

Yamashita, J. (1994). An analysis of relative clauses in the Lancaster/IBM spoken English corpus. *English Studies, 75*(1), 73–84.

Yang, J. (2020). *Using contextualized materials to teach English grammar.* [Master's thesis, The Faculty of the School of Education, University of San Francisco]

Authors

Yifan Feng

Master's Degree Student, Faculty of Liberal Arts, Mahidol University

E-mail: frankiefyf@gmail.com

Assistant Professor Suthathip Thirakunkovit, Ph.D.

Faculty of Liberal Arts, Mahidol University

E-mail: suthathip.thi@mahidol.edu