Studies in Linguistics and Literature
ISSN 2573-6434 (Print) ISSN 2573-6426 (Online)
Vol. 9, No. 4, 2025
www.scholink.org/ojs/index.php/sll

Original Paper

Variation in the Use of *Who* Relative Clauses by L2 English Writers: A Corpus-Based Study of TOEFL Essays

Zhupeng Li[1*]

[1] University of Connecticut, Storrs, United States

[*] Zhupeng Li, University of Connecticut, Storrs, United States

Received: October 27, 2025 Accepted: November 15, 2025 Online Published: November 30, 2025

doi:10.22158/sll.v9n4p81 URL: http://dx.doi.org/10.22158/sll.v9n4p81

Abstract

This study investigates how English learners at different proficiency levels use who relative clauses in their writing on the Test of English as a Foreign Language (TOEFL). Using the ETS Corpus of Non-Native Written English, the study examines 1,100 essays written by speakers of 11 first languages across low, medium, and high proficiency levels. Quantitative analyses identified the normalized frequencies of who relative clauses, while structural analyses categorized their syntactic types (OS, OO, SS, and SO). Results show that low-proficiency learners produced the highest normalized frequency of who relative clauses, whereas high-proficiency learners produced the fewest. Across all proficiency levels, OS and SS were the dominant types, confirming Keenan's (1975) relativized subject accessibility hierarchy. High-proficiency learners showed a greater proportion of SS types, reflecting increased syntactic complexity. Findings suggest that learners' proficiency levels influence the distribution and complexity of who relative clauses. The study concludes with pedagogical implications for sequencing grammar instruction and promoting syntactic development in EFL writing.

Keywords

relative clauses, who pronoun, TOEFL corpus, corpus linguistics, syntactic complexity, English proficiency

1. Introduction

Test of English as a Foreign Language (TOEFL) is a standardized test to measure the English language ability of non-native speakers who want to enroll in English-speaking universities. In looking at different proficiency levels, test takers' writing of the independent writing task has been collected to create the ETS Corpus of Non-Native Written English (also known as the TOEFL corpus) with the label of low, medium, and high (Blanchard et al., 2014). The complexity of test takers' syntactic performance is a criterion to judge their proficiency level. Among a great deal of linguistic structures, the relative clause is one of the major grammatical constructions in English language learning. This study investigates how TOEFL test takers across different proficiency levels use relative clauses, focusing on one relative pronoun—*who*. There are four different types of relative clauses, which are examined across TOEFL test takers in this study. The results can be used to understand the production frequency in terms of the four types of relative clauses for test takers in different proficiency levels.

2. Literature Review

According to Celce-Murcia and Larsen-Freeman (1999), a relative clause is "a type of complex post-nominal adjective modifier used in both written and spoken English" which is also known as a restrictive or adjective clause. Relative clauses provide a good source for investigating the underlying rules and processes that L2 learners use to process complex sentences (Abdolmanafi & Rahmani, 2012).

Language researchers, such as Kuno (1975) and Keenan (1975), proposed four types of relative clauses (i.e., OS, OO, SS, and SO), shown in Table 1. A few different natural sequencings for acquiring the four types of relative clauses are suggested. From the perspective of perceptual difficulty, Kuno (1975) argued that sentences with center embedding are perceptually harder for L2 learners to process than sentences with right branching relative clauses, indicating that OS and OO types should be easier than SS and SO types. From the perspective of relativized subject accessibility, Keenan (1975) hypothesized that relativized subjects are more accessible than relativized objected, suggesting that SS and OS types should be easier than SO and OO types. Finally, from the view of parallel function, Sheldon (1974) claimed that relative clauses which has the parallel functions as the head noun are much easier for L2 learners, so SS and OO should be easier than SO and OS types.

Table 1. Four Types of Relative Clauses

Type	Definition	Example
OS	The head noun is the object of the main clause, and the relative pronoun is the subject of the relative clause.	I know the student who got an A.
OO	The head noun is the object of the main clause, and the relative pronoun is the object of the relative clause.	I know the woman who (m) you are looking for.
SS	The head noun is the subject of the main clause, and the relative pronoun is the subject of the relative clause.	The student who got an A is a friend of mine.
SO	The head noun is the subject of the main clause, and the relative pronoun is the object of the relative clause.	The student whom you have talked to got an A.

A few studies have investigated the sequencings for acquiring the four types of relative clauses. Ioup and Kruse (1977) analyzed eighty-seven Chinese, Japanese, Arabic, Persian, and Spanish L2 learners' performance in grammaticality judgements, and their number of total errors revealed the order of acquisition of sentence types: OS > OO > SO >SS. Schumann (1980) investigated five Spanish and Italian L2 learners to determine the sequence of relative clause acquisition, and the frequency of production was OS > OO > SS >SO. Wong (1991) analyzed one hundred and seventy English compositions by L2 learners in a Hong Kong secondary school, and he found that the frequency of production or the acquisition sequence of relative clauses by these ESL learners was OS > OO > SS >SO. Even though some nuances are found among the findings, they confirm Kuno's hypothesis of avoidance of embedding.

3. Research Questions

This study aims to investigate whether there is a difference in the frequency of using *who* relative clauses across different proficiency levels, and whether, among four types of relative clauses, there is an order of difficulty or accessibility. A simple frequency count of attempts to use a particular type of relative clause can imply that this type is more easily accessible than other types not chosen for use as frequently.

To this end, the paper aims to answer the following research questions:

1) How does the use of *who* relative clauses differ across the three proficiency levels?

2) What is the rank order of production across the three proficiency levels among the four types of relative clauses (i.e., OS, OO, SS, and SO)?

4. Method

The corpus used in this study is the TOEFL corpus with low, medium, and high proficiency levels. It is comprised of 1100 English essays written by speakers of 11 non-English native languages (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish). The essays are presented in original raw forms and presented in UTF-8 formatted text files. The number of tokens in each level is 2,599,620 tokens in low, 18,991,261 tokens in medium, and 14,572,763 tokens in high.

The corpus was tagged for part-of-speech, which was used to search for *who* relative clauses by a regular expression. The study was conducted using TagAnt (2.4.0) and AntConc (3.5.9), with the former converting the original corpus to the tagged one and the latter searching for the target structure.

The target structure in this study is *who* relative clauses. Based on the TreeTagger Tag Set (Anthony, n.d.), a regular expression (RegEx) pattern was created to identify instances where *who* functions as a relative pronoun rather than as an interrogative. The specific RegEx pattern used was NN who_WP, which captures *who* clauses immediately following a noun phrase tag (NN). This method ensures that only *who* relative clauses are included in the frequency counts, minimizing unrelated occurrences of *who* in interrogative or other contexts. Both quantitative and structural analyses were undertaken in this study.

The tagged TOEFL corpus for low, medium, and high proficiency levels was loaded into AntConc (3.5.9). A regular expression search was conducted using the pattern NN who_WP to extract *who* relative clauses following noun tags. The resulting frequency counts were normalized per million words for cross-level comparison.

5. Results

5.1 Quantitative Analysis

Table 2 presents both the raw and normalized frequencies of *who* relative clauses across the three proficiency levels.

Table 2. Raw and Normalized Frequency of *Who* Relative Clauses

Proficiency	Raw Frequency	Normalized Frequency (per million words)
Low	236	907
Medium	1,566	824
High	1,132	773

Because the corpus sizes differ across proficiency levels, normalized frequencies were used for interpretation. The results indicate that low-proficiency test takers produced who relative clauses most frequently, while high-proficiency test takers used them least often. This suggests that as learners' proficiency increases, they rely less on this specific type of relative clause, possibly due to greater syntactic diversity in advanced writing.

5.2 Structural Analysis

The analysis focused on the four types of who relative clauses produced by test takers—OS, OO, SS, and SO. To enable comparison across proficiency levels while keeping the sample balanced, the first 200 instances of *who* relative clauses were analyzed from each proficiency level (low, medium, and high). The distribution of each type across levels is shown in Table 3.

Table 3. Frequency of Four Types of *Who* Relative Clauses across Proficiency Levels

Type	Low	Medium	Hight
OS (object–subject)	111	125	90
OO (object–object)	5	5	3
SS (subject–subject)	82	69	105
SO (subject–object)	2	1	2

Quantitatively, OS and SS were the most frequently produced clause types across all proficiency levels, whereas OO and SO were rarely used. Low- and medium-proficiency test takers displayed the same rank order of production (OS > SS > OO > SO), while high-proficiency test takers exhibited a different pattern (SS > OS > OO > SO).

In terms of structure, low-level learners often used OS clauses in simple descriptive sentences (e.g., *I know the man who lives next door*), suggesting reliance on familiar subject–verb patterns. Medium-level learners began to produce more SS clauses but occasionally showed structural errors (e.g., *The student who you told is my friend*). High-level learners used more SS clauses in complex noun phrases and embedded contexts (e.g., *The teacher who encouraged me to study abroad inspired my research*), demonstrating greater syntactic flexibility.

These findings indicate developmental progression from reliance on straightforward right-branching OS clauses toward more complex subject-relative constructions as proficiency increases.

6. Discussion

The first research question examined how the use of *who* relative clauses varies across proficiency levels. Quantitative results revealed that low-proficiency test takers produced the highest frequency of *who* relative clauses, whereas high-proficiency test takers produced the fewest. This pattern suggests that lower-proficiency learners rely more on familiar and formulaic syntactic structures such as *who* relative clauses to organize information. In contrast, higher-proficiency learners appear to diversify their syntactic repertoire by employing other relative pronouns (e.g., *which, that*) or alternative complex constructions, such as prepositional phrases or reduced clauses, to achieve cohesion and variation in their writing.

The second research question investigated differences in the distribution of four relative clause types (OS, OO, SS, SO) across proficiency levels. The results showed that low- and medium-proficiency learners followed the same production order (OS > SS > OO > SO), while high-proficiency learners displayed a different pattern (SS > OS > OO > SO). Across all proficiency levels, OS and SS clauses predominated, whereas OO and SO clauses occurred infrequently. This finding confirms Keenan's (1975) relativized subject accessibility hierarchy, indicating that subject-relative clauses are cognitively easier to process and produce than object-relative clauses.

A more detailed comparison between OS and SS clauses demonstrates a developmental shift. Low- and medium-proficiency learners produced more OS clauses, which typically involve right-branching, perceptually simpler structures (e.g., *I know the student who lives next door*). High-proficiency learners, however, produced more SS clauses, which require greater syntactic control and often occur in more integrated noun phrases (e.g., *The teacher who inspired me encouraged my study abroad*). This pattern aligns with Kuno's (1975) perceptual difficulty hypothesis, which argues that learners with higher syntactic competence can process center-embedded or more complex constructions more easily. The transition from OS-dominant to SS-dominant usage thus reflects increasing grammatical sophistication and structural flexibility among advanced writers.

7. Pedagogical Implications

The findings of this study provide several implications for language teaching, particularly for grammar instruction and writing development in EFL and ESL contexts.

7.1 Sequencing Grammar Instruction

The frequency and developmental patterns observed support introducing relative clauses in a staged manner – beginning with subject-relative types (SS and OS) before progressing to object-relative types (OO and SO). Teachers can design tasks that build learners' confidence with simpler clause structures before gradually introducing more complex forms, and eventually the complex tasks can facilitate language development in terms of language complexity, accuracy, and fluency (Li, 2024).

7.2 Promoting Structural Variety in Writing

Since lower-proficiency learners tend to overuse *who* clauses, instructors can provide targeted feedback and corpus-based materials to help students diversify their syntactic patterns. For example, contrastive tasks highlighting how advanced writers use which or that clauses can raise learners' awareness of alternative relative structures.

7.3 Integrating Corpus Tools for Data-driven Learning

Tools such as AntConc and learner corpora (e.g., the TOEFL corpus) can be used in classroom activities to help learners notice authentic examples of relative clause usage. This approach promotes data-driven learning, which encourages learners to discover grammatical patterns and variation independently rather than memorizing rules in isolation. Language teachers and educators should be flexible to choose the most appropriate approach depending on their teaching context (Li, 2025).

7.4 Developing Metalinguistic Awareness with Authentic Examples

Analyzing authentic examples of relative clauses can help students understand not only the structural formation of clauses but also the cognitive factors that make some types easier or harder to process. According to Li (2023), language classes should be designed to engage learners into authentic grammar-learning tasks. Authentic activities on syntactic complexity can foster deeper awareness of how grammar supports meaning and rhetorical effect in writing.

8. Conclusion

This study examined the use of *who* relative clauses across three proficiency levels in the TOEFL learner corpus, revealing both quantitative and structural developmental differences. Lower-proficiency learners relied more heavily on *who* relative clauses, whereas higher-proficiency learners used them less frequently but with greater syntactic complexity. The shift from OS- to SS-dominant patterns across proficiency levels supports both Keenan's accessibility hierarchy and Kuno's perceptual difficulty hypothesis, illustrating the cognitive and structural progression in learners' interlanguage development.

Pedagogically, these findings highlight the value of corpus-based approaches for understanding grammatical development and designing data-informed instruction. By combining frequency-based insights with pedagogical applications, this study underscores how corpus linguistics can inform effective grammar teaching and bridge the gap between linguistic theory and classroom practice.

References

Abdolmanafi, S & Rahmani, Ali. (2012). An Investigation of the Learnability of Relative Clauses by EFL Learners. *World Journal of English Language*, *2*(3), 1-9.

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University.

Anthony, L. (2022). TagAnt (Version 2.4.0) [Computer Software]. Tokyo, Japan: Waseda University.

Anthony, L. (n.d.). *TreeTagger Tag Set*. https://www.laurenceanthony.net/software/tagant/resources/treetagger_tagset.pdf

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2014). *ETS Corpus of Non-Native Written English.* https://catalog.ldc.upenn.edu/LDC2014T06

Celce-Murcia, M. ＆ Larsen-Freeman, D. (1999). *The Grammar book: An ESL/EFL teacher's course* (2nd ed.). Massachusetts: Heinle ＆ Heinle.

Ioup, G., & Kruse, A. (1977). Interference versus structural complexity in second language acquisition: Language universals as a basis for natural sequencing. In R.D. Brown, C.A. Yorio, & R.R. Crymes (Eds.), On *TESOL* 77: *Teaching and learning English as a second language: Trends in Research and Practice*. Washington, D.C.: TESOL.

Keenan, E.L. (1975). Variation in universal grammar. In Fasold and R. Shuy (Eds.), *Analysing variation in language*. Washington DC: Georgetown University Press.

Kuno, S. (1975). The position of relative clauses and conjunctions. *Linguistic Inquiry*, *5*, 117-136.

Li, Z. (2023). Effects of focused and unfocused tasks on L2 learners' grammatical development: A selective review of literature. *Journal of Foreign Language Education and Technology*, *8*(1), 1-10.

Li, Z. (2024). The Effects of Task Complexity on Second Language Learners' Pragmatic Knowledge Development: A Review of Literature. *Iris Journal of Educational Research*, *3*(2), 1-5.

Li, Z. (2025). [Review of the book *Common ground: Second language acquisition theory goes to the classroom* by Florencia G. Henshaw & Maris D. Hawkins (2022)]. *Teaching English as a Second Language Electronic Journal (TESL-EJ)*, *28*(4). https://doi.org/10.55593/ej.28112r2

Schachter, J. (1974). "An error in error analysis", *Language Learning*, *24*, 205-214.

Schumann, L. (1980). The acquisition of English relative clauses by second language learners. In R. C. Scarcella, & S. D. Krashen (Eds.), *Research in Second Language Acquisition*. Rowley, Mass: Newbury House.

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

Wong, J. (1991). Learnability of relative clauses: A Hong Kong Case. *Working Papers of the Department of English*, *3*(1), 108-117.