To Cleave or Not to Cleave: Distributional Frequencies of Cleft Structures in Research Articles, Textbooks, and PhD Dissertations

Amir Mohammad Yousefi [1] , Rajab Esfandiari [2*]

[1] *MA in English Language Teaching, Department of English Language Teaching, Imam Khomeini International University, Qazvin, Iran*
[2] *Associate Professor of English Language Teaching, Department of English Language Teaching, Imam Khomeini International University, Qazvin, Iran*

Abstract: Although researchers have analyzed the formal, syntactic, and functional behavior of cleft sentences in the English language across various genres, their distributional frequencies have received very little attention in academic research genres. Therefore, drawing on a 20,389,297-word corpus including 1,521 research articles (RAs), 116 PhD dissertations, and 48 textbooks, in applied linguistics, this study followed Biber et al.'s (1999) and Collins' (2002) models to identify and analyze four major types of cleft structures in our corpus. Using a corpus-based research design, we drew on concordances to extract the target cleft structures. The computer program AntConc was used to identify instances of cleft sentences, and the statistical significance of the findings was evaluated through separate chi-square tests. The results of frequency analysis showed varying degrees of these grammatical structures across the three research genres, with textbooks including the highest number of clefts, followed by dissertations, and RAs. The results promise pedagogical implications for non-native English student writers to familiarize themselves with conventions in academic research genre writing for publication purposes. The findings suggest that academic research genre writing affects the frequency and use of these complex grammatical structures.

Keywords: Academic Writing, Cleft Constructions, Research Articles.

_____

* **Corresponding Author.**

Authors' Email Address:
[1] **Amir Mohammad Yousefi** (yousefi.amir95@gmail.com), [2] **Rajab Esfandiari** (esfandiari@hum.ikiu.ac.ir)

Introduction

The analysis of language patterns in academic genres has recently attracted the attention of numerous researchers. As Hyland (2023) has argued, academics display a considerable amount of dexterity in representing their disciplinary identity, which helps them demonstrate their loyalty to their discipline and competent participation in a disciplinary community. Academic writing requires a more refined style of writing. Academics tend to use a more elaborate means of developing, supporting, or countering arguments, and leading the readership through logical steps to a conclusion (Conrad, 2000). Developing competence in order to use language appropriately is crucial because claims to knowledge are made through language (Henderson et al., 1993). Non-native English speakers do not necessarily use the same patterns as native-English speakers do, and even if they use the same patterns, they might use them functionally differently. According to Hyland (2019), this may be due to the cultural differences between native and non-native English speakers, which makes the comprehension of texts difficult for native English speakers. Clearly, it is fundamental for non-native English speakers to familiarize themselves with the structural patterns that native English speakers employ in English for Academic Purposes (EAP).

Academics genres, including research articles (RAs), textbooks, and dissertations, in EAP are one of the most important areas of academic writing in academic settings (Matikainen, 2024; Zotzmann & Sheldrake, 2021) for students. As Hyland and Hamp-Lyons (2002) noted, students need to master English in order to succeed in their courses "through the medium of English in textbooks, lectures, study groups, and so on" (p. 2). Academic genres in EAP settings are to disseminate knowledge, help writers share ideas with one another, and enable researchers to publicize the most recent developments in various fields of study (Becher & Trowler, 2001).

Cleft Structures in the English Language

As a type of relative clause, a cleft sentence is a complex sentence, including a main clause and a dependent clause, which denotes a meaning that a simple sentence can express. Some researchers claim that the clefts in the English language, mainly it-clefts, originated in early Middle English (Ball, 1991, 1994), while others believe their origin can be traced back to Old English (Patten, 2012). In the following paragraphs, the views of Collins (2002), Patten (2012), and Biber et al. (1999) on cleft sentences will be examined in detail.

Theoretical Foundations of Clefting: Collins' (2022) **Functional View and** Biber et al.'s (1999) **Corpus-Based Analysis**

Cleft structures can be studied from several perspectives. Using a functional approach, Collins (2002) divides these structures into three categories: Clefts (also known as it-clefts, or basic clefts), pseudo clefts, and reverse pseudo clefts. In cleft structures, unlike simple sentences, the material is divided into distinct sections. The section that immediately follows the copula (be derivatives such as is) in the superordinate clause is often called '*focus*' or '*stressed item*'. The constituent which is introduced by the relative pronoun is known as '*presupposition*' in the literature. However, in order to avoid confusion and misunderstanding with the work of Halliday (1994) on the subject, Collins uses the terms '*highlighted element*' and '*relative clause*' to refer to focus and presupposition, respectively. According to Collins (2002), cleft structures are identifying constructions, which "*express a relationship of identity between the elements realized as the highlighted element and the relative clause*" (p. 2). Identifying constructions need to be distinguished from attributive constructions. Collins (2002) describes identification as the relationship between an entity and attributes that are ascribed to it. These attributes can be an indication of class membership, quality, role, or such characteristics. One difference between identifying and attributive construction is that the former is typically reversible.

It-clefts, also referred to as clefts or simple clefts, are one form of cleft structures. Prince (1978) gives the following general formula for it-clefts: "*It is/was Ci which/who(m)/that Ø S-Ci*" (S-Ci in the formula refers to Sentence minus Constituent) (p. 883). However, Collins argues that this formula needs to be modified. One reason for this modification is "*the superordinate clause may select for modality, aspect, and polarity, and may include a 'focusing adverb' (only, just, and so on) between the copula and the highlighted element*" (p. 34). Another reason is that when, where, in which, and for whom can be added to the possible wh-words. The last reason for changing the formula proposed by Prince (1978) is that in some instances of cleft sentences with experiential function, the highlighted element can be regarded as optional. Therefore, the following formula is proposed by Collins (2002) to account for all possible combinations allowed to occur with it-clefts: [It + (Modal) (NEG) (ADV) (have) [be (NEG)(ADV) (Ci)]+which/whom/who/that/when/where/Ø] S-Ci.

According to Collins (2002), pseudo clefts are comprised of three subclasses: wh-clefts, th-clefts, and all-clefts. Wh-clefts, as the name suggests, are relative clauses headed by a wh-item. Some of these wh-items are: what, who, where, when, why, or how. Similarly,

relative clauses headed by th-item such as *the*, are called th-clefts. Last but not least, nominal clauses that start with the word *all* are named all-clefts.

According to Biber et al. (1999), clefting is similar to dislocation, meaning that the information which is given in a single clause, is broken down into two clauses, each having its own verb. Two major types of cleft sentences are it-clefts and wh-clefts. Both types are used to bring particular elements into focus or show contrast. The extra focus usually appears early in it-clefts and late in wh-clefts to help with cohesion and distribution (Biber et al., 1999).

One of the most common forms of cleft sentences is It-Clefts. An it-cleft sentence consists of the pronoun *it*, followed by a conjugated form of the verb *be*, which is optionally accompanied by the negator *not* or an adverb such as *only*. Then, the focused element that follows can be a noun phrase, a prepositional phrase, an adverb phrase, or an adverbial clause. Lastly, a relative-like dependent clause introduced by that, who/which, or a zero relativizer, whose last element receives normal end-focus will follow (Biber et al., 1999).

Biber et al. (1999) investigated wh-clefts as well. This form of cleft sentence consists of a clause introduced by a wh-word, usually *what*. Next, a form of the verb *be* will follow. The specially focused element which can be a noun phrase, an infinitive clause, or a finite nominal clause will be next. The point of focus in wh-clefts is typically located at the end (Collins, 2002).

Another form of wh-clefts is the reversed wh-cleft. Some of these reversed versions look exactly like ordinary wh-clefts, but the position of the wh-clause in relation to the focused element is different (Biber, et. al., 1999) (e.g., you see **a weekend flight** is what you want.). According to Biber et al. (1999), a very common type of structure contains a demonstrative pronoun, usually *that*, followed by a form of *be* plus a dependent clause introduced by a wh-word. It should be noted that although these structures are not reversible, they are structurally related to the reversed wh-clefts. Wh-clefts that open with a reference to the preceding texts are called demonstrative wh-clefts (Calude, 2017).

Researchers have studied different genres of academic writing in different disciplines over the years. Some researchers have looked into the genre of research articles, studying different patterns and structures such as how criticality is expressed in literature review in research article introductions (Bruce, 2014), and how the niche is established in the introduction sections (Moghaddasi & Graves, 2017). Other researchers have examined the genre of PhD dissertations, studying the functions of modifiers in shaping dynamic relationships in dissertation defenses (Lin, 2017), how students express their stance in the

acknowledgment section (Chan, 2015), how research questions are formulated (Lim, 2014), and how non-native students and their advisors use self-reports in dissertation writing (Dong, 1998). Finally, some researchers have also studied the genre of textbooks and academic writing. They have investigated the use of imperatives in academic writing by students (Swales & Post, 2018), and how writers use conditional clauses to mold interpersonal relations in written academic discourse (Warchał, 2010). However, researchers have paid very little attention to clefting in RAs, textbooks, and PhD dissertations. The present study, accordingly, aimed to fill this gap and the following three research questions were analyzed.

1) Are there any differences in the frequency of cleft structures in research articles between English-speaking writers and Iranian writers?

2) Are there any differences in the frequency of cleft structures in PhD dissertations?

3) Are there any differences in the frequency of cleft structures in textbooks?

Review of the Literature

The distribution of cleft constructions in conversation, fiction, news, and academic prose has been examined. Collins (2002), for example, examined cleft constructions using the following corpora: The London-Lund Corpus of Spoken English (LL) and Lancaster-Oslo-Bergen (LOB). The LOB corpus includes genres such as informative prose (i.e., press, skills, trades and hobbies, popular lore, belles, letters, biography, essays, government documents, reports, catalogs, learned and scientific writings) and imaginative prose (i.e., general fiction, mystery and detective fiction, science fiction, adventure and western fiction, romance and love story, and humor). The findings showed more frequent uses of pseudo-clefts than clefts in speech. By contrast, clefts were more frequently used in writing. Interestingly, reversed pseudo-clefts were more frequently used in both speech and writing. According to Collins (2002), the more informal the situation, or the more familiar interlocutors are with each other, the more reversed pseudo-clefts are used. However, when the interlocutors do not know each other very well, as in monologues, public dialogues, and telephone conversations, they might use pseudo-clefts in order to be more formal and distant. The same patterns can be observed in the written genres as well. The it-cleft structures significantly outnumber pseudo-clefts and reversed pseudo-clefts in informative prose such as reports, scientific writing, biographies, and essays. However, in imaginative prose such as general fiction, romance and love stories, and humor, pseudo-clefts and reversed pseudo-clefts were utilized significantly more than

clefts. The similarities between imaginative prose and speech, such as dialogues and conversations, might explain the reason for the abundance of pseudo-clefts and reversed pseudo-clefts in this genre.

Biber et al. (1999) also examined cleft structures using corpus data. According their findings, it-clefts are relatively common in all registers, but they are most frequent in academic prose. Ordinary wh-clefts are most frequent in conversation. Meanwhile, reversed wh-clefts are infrequent in all registers. Lastly, the frequency of demonstrative wh-clefts differs by register. While common in conversation, they are rare in academic prose.

Hasselgård (2014) investigated it-clefts in L1 and L2 academic writing. She studied it-clefts in Norwegian learners' linguistic papers and argumentative essays from several aspects namely: frequency of use, grammatical features and grammatical context, and discourse functions of clefts. The comparison between different corpora in the study showed significant underuse in cleft structures when The International Corpus of Learner English (ICLE-NO) was compared to The Louvain Corpus of Native English Essays (LOCNESS) and The Varieties of English for Specific Purposes database learner corpus (VESPA-NO) to The British Academic Written English (BAWE). However, when BAWE was compared to BNC, there was a significant overuse of BAWE.

In addition to the frequency of cleft structures, researchers have also investigated their functions. As discussed earlier, all cleft structures focus on part of the sentence to highlight prominence (Goldberg, 2006; Quirk, et. al., 1985). However, there are important differences between the types of cleft structures. It-clefts carry a higher load of information compared to the pseudo-clefts and reversed pseudo-clefts (Collins, 2002). The given information is expressed by the focused element of the It-cleft, which is not infrequently a pronoun or some other form. The early position of the focused element in it-clefts is both suitable for expressing a connection with the preceding text and a contrast (Calude, 2017; Hasselgård, 2014). It-clefts are also less personal in nature, which allows writers to distance themselves from the opinions they provide (Collins, 2002)

Pseudo-clefts, according to Collins (2002), are characterized by explicit specification and shared knowledge that is utilized best when used in speech. The positioning of the focused element in the ordinary wh-clefts is in agreement with the information principle (Collins, 2002). According to Biber et al. (1999), expressing new information is the job of the focused element in the ordinary wh-clefts. Pseudo-clefts help the speaker specify the background, or shared knowledge, that the addressee is supposed to have before delivering the new information (Collins, 2002). Similarly, reversed pseudo-clefts have a structure that is

suited for the dynamic organization of spoken language. In the reversed wh-clefts and demonstrative wh-clefts, the focused element is typically context-dependent (Calude, 2017). According to Collins (2002), reversed pseudo-clefts serve a summative role. There might be two reasons for this: the first reason is that since reversed pseudo-clefts usually occur at the end of the sentence, they signal low informational content. The second reason is that reversed pseudo-clefts operate as internal referencing devices. Collins (2002) adds that reversed pseudo-clefts provide little to no new information. By doing so, reversed pseudo-clefts round off the paragraph and, at the same time, set up the next paragraphs. These might be the reasons why reversed pseudo-clefts are believed to be more suited for stage-ending roles.

Cleft constructions are found in both conversation and formal written registers (Biber, et. al., 1999; Calude, 2007, 2017; Goldberg, 2006; Quirk et al., 1985). It-clefts are especially common in academic prose due to the fact that they allow very precise statements to be made (Biber et. al., 1999). Wh-clefts are used in conversations mostly because of the low information content that wh-clefts carry (Collins, 2002; Biber et. al., 1999). Like other fronted elements, reversed wh-clefts are infrequent (Biber, et. al., 1999). Finally, the demonstrative wh-cleft is used in informal registers like conversations (Calude, 2017). According to Biber et al. (1999), "This is supported by the behavior of constructions opening with this and that; the more formal this is, in fact, the preferred form in academic prose" (p. 963).

Methodology

Corpus Development, Corpus Size, and Corpus Constituents

The corpus of the present study includes PhD dissertations, research articles (RAs), and textbooks in applied linguistics. In order to make the data as representative as possible, the corpus sizes were set to a minimum of five million words per genre. PhD Dissertations written by English-speaking writers were downloaded from http://www.proquest.com. The dissertations were published during the 2010-2018 period. The eight-year interval was selected to gain an understanding of the recent practices in dissertation writing. With a total of 116 dissertations, the number of words in the dissertations corpus reached a total of 5, 032, 129 words.

RAs written by English writers were selected from available journals in applied linguistics. To find the journals in applied linguistics, we followed the guidelines and the procedures in Hashemi and Babaii (2013). These sources were the lists of professional journals published by The Modern Language Journal (Weber & Campbell, 2004), Egbert's

(2007) evaluation of applied linguistics journals, Jung's (2004) examination of the frequency of appearance of ELT journals selected for presentation in *Language Teaching* between 1996 and 2002, and Lazaraton's sample (2000). The format of an article may, or may not, affect the choice of cleft structures used by writers. Therefore, following the model introduced by Swales (1990), journals not following the IMRD format (Introduction, Methods, Results, and Discussion) were excluded from the corpus.

There are many factors affecting a journal's quality. Journals are often judged by citation analysis, rejection rate, time of publication, and expert opinion (Egbert, 2007). However, citation analysis has remained one of the most common means of determining a journal's impact factor (Brumback, 2009; Leydesdorff & Opthof, 2010; Weiner, 2001). The writing guidelines that each journal imposes on the authors may affect the style and format of articles as well. That is why in the present study, limiting the scope of research, and giving more validity to the data, journals with a 5-year impact factor higher than 1.8 were selected. It must be noted that initially, a 5-year impact factor of 2 was selected as an average impact factor criterion for journal selection. However, since not many journals in applied linguistics had such a high 5-year impact factor, in order to reach the five-million-word goal, it was lowered to 1.8. Articles written by English writers published in 2010-2018 were selected. Articles with multiple authors were included only if all the authors were native English speakers; otherwise, they were excluded from the corpus. At first, a sample of one hundred articles was randomly downloaded and the average words per article was calculated. From that average, the total number of articles required to be downloaded from each volume became apparent. This method proved to be problematic since the number of articles written by English writers was not equal in different journals. Thus, in order to reach the five-million size, all of the RAs written by English writers that fell in between the 2010-2018-year interval were downloaded. The details of each journal are provided in Table 1.

No agreed-upon consensus is found in the published literature to help researchers identify published native English speakers. However, we drew on the following two procedures to include in the native corpus the native English-speaking writers. First, we built on Wood (2001), who claimed that when the authors' names sound English and the authors are affiliated with an English institute, they are regarded as native English speakers. Second, we compared the authors' names against the 900-3000 list of English names ((covering 88.6% and 88.5% of the male and female population, respectively) by Lu and Deng (2019) collected using the 1990 US Census name files. Although we used these two procedures, we

might have included in our corpus the authors who may not necessarily have been native English speakers, a limitation that has to be considered in the present study.

Table 1. Name and Information of International Journals

Journal	Issues	5-year Impact Factor	Word Count
Applied Linguistics	38	3.899	577,750
English for Specific Purposes	36	1.829	354,580
Journal of English for Academic Purposes	40	2.093	466,129
Journal of Second Language Writing	36	3.146	426,389
Language Teaching Research	44	2.536	522,707
Studies in Second Language Acquisition	36	3.146	484,111
TESOL Quarterly	36	2.704	637,456
The Modern Language Journal	36	2.578	956,434
Written Communication	36	2.675	733,199
Total			5,158,755

RAs written by Iranian writers from 2010-2018 were selected from available Iranian journals in applied linguistics. Since the number of available RAs in Iranian journals was much fewer than their native counterparts, all of the articles written by Iranian researchers in the journals were downloaded. Unfortunately, unlike international journals, Iranian journals are not rated by any factors yet. Iranian journals that had already received a license from the Ministry of Science, Research, and Technology of the Islamic Republic of Iran were included in the corpus. Some journals such as Journal of Modern Research in English Language Studies had just recently acquired the proper license from the ministry and, therefore, did not have enough articles. Such journals were removed from the corpus. Name and details of the journals used in the current study are provided in Table 2.

Table 2. Name and Information of Iranian Journals

Journal	Issues	Word Count
Applied Research on English Language	25	468,635
Iranian Journal of Applied Linguistics	18	709,464
Iranian Journal of Language Teaching Research	11	275,914
Issues in Language Teaching	11	402,856
Journal of English Language Teaching and Learning	18	624,386
Journal of Research in Applied Linguistics	17	700,518
Journal of Teaching Language Skills	33	1,245,467
Teaching English Language	18	714,371
		5,141,611

Finally, textbooks written by English-speaking writers from 2010-2018 were selected. The textbooks were all in the field of applied linguistics and were targeted toward the BA and MA students. In order to select the textbooks, the syllabi of some universities in Iran (such as Imam Khomeini International University in Qazvin and the University of Tehran in Tehran) were referred to. The total number of textbooks in the corpus reached 48, with 5,056,802 words. The books included in the present study covered a variety of subjects. Subjects such as applied linguistics, corpus linguistics, discourse analysis, ESP, grammar, pragmatics, research, syllabus design, teaching, testing, vocabulary, and writing are covered by the books included in the corpus. The number of words in each book ranged from 14,249 to 445,557 words.

Compiling the corpus for the present study proved to be a cumbersome task. Luckily, the number of words in each subcorpus surpassed the minimum of one million words set by O'Keeffe and McCarthy (2010). Gathering a representative dataset that includes the full range of variability in a population (Biber, 1993) was the goal of the present researcher. A summary of each corpus is provided in Table 3.

Table 3. Information on the Research Genres, the files, and Length of the Subcorpora

Name of the corpus	Number of files	Number of Words
PhD Dissertation-English	116	5,032,129
RAs-English	636	5,158,755
RAs-Iranian	885	5,141,611
Textbooks-English	48	5,056,802
Total	1685	20,389,297

Instrumentation

Three computer programs were used in the present study. First, the computer program Antconc (Anthony, 2018) version 3.5.7 was used to analyze the corpus data in this study. This software is available for download from: http://www.laurenceanthony.net/software/antconc/. To convert and edit the PDF files into Microsoft Word documents, Adobe Acrobat DC was used. Lastly, IBM SPSS version 25 was used for the statistical analysis.

To identify the cleft structures, Biber et al.'s (1999) framework was adopted in the present study. They divided cleft structures into two major categories of it-clefts and wh-clefts. The following formula is put forward by Biber et al. for it-clefts: *it + verb be + optional negator/adverb + focused element + dependent clause introduced by that/who/which/zero*.

Wh-clefts are formulated as a clause introduced by a *wh-word + verb be + focused element*. In reversed wh-clefts, the focus element is shifted to the beginning of the sentence. Biber et al.'s formula for cleft structures is clear and straightforward, which makes it a suitable option for using in computer programs such as AntConc. Biber et al., however, did not take some of the pseudo-clefts into consideration, which motivated the presented researchers to draw on Collins' (2002) framework as well.

Collins (2002) has focused mostly on pseudo-clefts in his study. Other than wh-clefts and reverse wh-clefts, Collins examined all-clefts and th-clefts as well. Collins put forward the following formula for pseudo-clefts: [what/who/where/when/why/how/the + (ADV) + thing/one/place/time/reason/way + (that/PP/which/why/when/where/who/whom/zero)] + [S-C_i]. The comprehensive formula for pseudo-clefts aside, Collins also introduced some tests to identify true pseudo-clefts such as the reversibility test and the uncleavability test. These tests were used to see if the pseudo-clefts were reversible and could have uncleft counterparts. If a cleft sentence passed the two tests, it could be considered a genuine cleft structure.

Biber et al.'s formula for it-clefts and wh-clefts was used to identify the cleft structures in the present study. To have a comprehensive list of pseudo-clefts in the present study, all-clefts and that-clefts (referred to as demonstrative wh-clefts by Biber et al.), studied by Collins were added to pseudo-clefts introduced by Biber et al. (1999). The context words such as *all, one, place, reason, that, thing, time,* and *way* introduced by Collins (2002) were used to narrow down the concordance lines provided in AntConc. The uncleavability and

reversibility tests were used to judge the genuineness of pseudo-clefts found in the present study.

Procedure and Identification of Cleft Structures

The process of corpus development began by collecting data in the field of applied linguistics from three sources of RAs, textbooks, and PhD dissertations. All of the downloaded material were converted from PDF files into Microsoft Word documents using the Adobe Acrobat DC software. After the conversion, charts, diagrams, tables, reference lists, acknowledgments, and forewords were deleted from the texts. The remaining texts were saved as .TXT files, and their words were counted. In the case of the files that were difficult or problematic to convert, the whole article was either replaced or reconstructed by the researchers and double-checked afterward.

Known patterns and keywords of cleft structures were introduced to the AntConc software. The following patterns were searched to narrow down the possible cleft sentences: *it is/was/were, what I/he/she/they, all I/he/she/they, that is/was/were*. The following context words, which were used by Collins (2002), were included in the search to further narrow down the results: *all, one, place, reason, that, thing, time, way, are, is, was, were, he, she, I, we, researcher*. The exact search terms, context words, and settings used for identifying each cleft construction are provided in the appendix. The concordance lines provided by the software were analyzed by the researchers, and clefts were separated from the non-cleft structures.

Because of the work of Collins (2002), pseudo-clefts were the least problematic to identify. Both reversibility and uncleavability tests were used to make sure the selected pseudo-clefts were true clefts. However, it-clefts were a bigger challenge to identify. Since there is no agreement between researchers, it-clefts are difficult to identify and classify. In order to have a comprehensive representation and data, patterns of it-clefts introduced by both Biber et al. (1999) and Collins (2002), and instances of it-clefts found in other available corpora were taken into consideration. This method generated more comprehensive results which not only included the common forms of cleft structures, but also other varieties of it-clefts and pseudo-clefts. Some traditional grammar books (Larsen-Freeman, 1993; Thomson & Martinet, 1986), for example, hold the belief that it-clefts can only have NPs as focal heads and other varieties are not true clefts. Some contemporary grammar books (Patten, 2012) still regard NPs as the most common heads for cleft structures and have mixed views on other types of focal heads. Such cleft structures, especially in this case it-clefts, which deviated

slightly from the known and agreed-upon formula, were included in the corpus based on the following two criteria. First, they were included in the present study if they had functions such as contradiction-making, summarizing, or focus-bringing as most clefts do. The second criterion was the uncleavability test. If the non-cleft counterpart did not result in an ungrammatical or ambiguous sentence, they were included in the study.

After the identification process was over, the concordance lines were analyzed by the two researchers. The initial Pearson correlation between the two experts showed an acceptable level of agreement (r = .86). The concordance lines were examined again and the problematic cleft structures were removed. Afterward, the results were sent to the third expert for a second time. New comments provided by the third expert were taken into consideration as well. The intraclass reliability test was used to get a more conservative and accurate measure of reliability (r = .94). The results of the intraclass correlation coefficient proved satisfactory, and the remaining differences were discussed and resolved.

Research Design of the Study

In the present study, the primary focus was to identify a series of linguistic elements in a specialized researcher-developed corpus. Therefore, the researchers built on a corpus-based design to develop the corpus using three distinct genres: RAs, textbooks, and dissertations. Using concordance lines in the corpus, we counted the number of times the cleft structures used in the corpus, examined the structural behavior of these elements, and compared the final outcomes across the three genres.

Results

Research Question 1: *Cleft Structures in RAs Between English-speaking Writers and Iranian Writers*

The first research question examined cleft structures between English-speaking and Iranian writers in RAs. Table 4 summarizes the descriptive statistics for cleft structures identified in the RAs written by both English-speaking and Iranian writers. As can be seen in Table 4, it-clefts were the most frequent types of cleft structures used by both English and Iranian writers in RAs. The next most frequent type of cleft structure found in RAs was wh-clefts. Finally, both all-clefts and that-clefts were the least frequent cleft structures in both English and Iranian written RAs.

Table 4. Descriptive Statistics of Cleft Structures in RAs

Clefts	writer	
	English-speaking	Iranian
All	6	8
It	169	95
That	7	8
Wh	19	10
Total	201	121

Table 5. Frequency of Cleft Structures in RAs

		All-cleft	It-cleft	That-cleft	Wh-cleft
Writer	English	Count	6	169	7	19
		Expected Count	8.7	164.8	9.4	18.1
		% within Writer	3.0%	84.1%	3.5%	9.5%
		Standardized Residual	-.9	.3	-.8	.2
		Adjusted Residual	-1.5	1.3	-1.3	.4
	Iranian	Count	8	95	8	10
		Expected Count	5.3	99.2	5.6	10.9
		% within Writer	6.6%	78.5%	6.6%	8.3%
		Standardized Residual	1.2	-.4	1.0	-.3
		Adjusted Residual	1.5	-1.3	1.3	-.4

To examine statistically significant differences, a 2 x 4 chi-square test was used. As Table 6 shows, no statistically significant differences were found in the number of cleft structures used by English and Iranian writers in RAs [χ^2 = 4.276, p = .322, Cramer's V = 0.115].

Table 6. Chi-square Test for RAs between English and Iranian Writers

	RAs
Pearson chi-Square	4.276
Df	3
Asymp.Sig	.322
Cramer's V	.115

The Second Research Question: *Cleft Structures in PhD Dissertations*

The second research question investigated the frequency of cleft structures by English speakers in PhD dissertations. Table 7 summarizes the descriptive statistics for cleft

structures identified in the dissertation corpus. Table 7 indicates that the most frequent cleft structure used by English writers in dissertations is it-clefts. The next most frequent cleft structure found in dissertations is wh-clefts. Both all-clefts and that-clefts were the least occurring cleft structures in the dissertations.

Table 7. Descriptive Statistics of Cleft Structures in Dissertations

	Dissertation Clefts
All	9
It	172
That	9
Wh	42
Total	232

Table 8. Frequency of Cleft Structures in Dissertations

Clefts	Observed N	Expected N	Residual
All	9	58.0	-49.0
It	172	58.0	114.0
That	9	58.0	-49.0
Wh	42	58.0	-16.0
Total	232		

To find out whether the number of clefts present in the dissertations written by English-speaking writers was significant or not, a one-way chi-square test was used (Table 9). Table 9 shows that the number of clefts in the dissertations corpus was statistically significant [$\chi^2 = 311.276$, $p = .000$]. It is noticeable in Table 8 that it-clefts contributed the most to the observed chi-square.

Table 9. Chi-square Test for Dissertations

	Dissertations
Pearson chi-Square	311.276
Df	3
Asymp.Sig	.000

The Third Research Question: *Cleft Structures in Textbooks*

The third research question examined the frequency of cleft structures in textbooks written by English-speaking writers. Table 10 summarizes the descriptive statistics for cleft structures identified in the textbook corpus. As can be seen in Table 10, the most frequent cleft structure used by English writers in the text books genre is it-clefts. The next most frequent cleft structure found in dissertations is wh-clefts. That-clefts take the third place in the frequency list, followed by all-clefts as the least frequent.

Table 10. Descriptive Statistics of Cleft Structures in Textbooks

	Textbook Clefts
All	9
It	372
That	10
Wh	87
Total	478

In order to answer the third research question regarding the presence of cleft structures in textbooks, a one-way chi-square test was used (Table 12). Table 12 indicates that the differences in the number of clefts in the textbooks corpus were statistically significant [χ^2 = 744.879, p = .000]. Table 11 shows that it-clefts were the main contributors to the observed chi-square.

Table 11. Frequency of Cleft Structures in Textbooks

Clefts	Observed N	Expected N	Residual
All	9	119.5	-110.5
It	372	119.5	252.5
That	10	119.5	-109.5
Wh	87	119.5	-32.5
Total	478		

Table 12. Chi-square Test for Textbooks

	Textbooks
Pearson chi-Square	744.879
Df	3
Asymp.Sig	.000

All the focal heads of cleft structures in the present study were checked to find out the different variations used by each group of writers. A variety of phrases was observed in RAs,

PhD dissertations, and textbooks, as can be seen in Table 13. NPs were the most frequently used phrase by all groups of writers.

Table 13. Phrases Used in the Cleft Structures

	Clefts	Phrases				
		Noun phrase (NP)	Adjective phrase (ADJP)	Adverb phrase (ADVP)	Prepositional phrase (PP)	Verb phrase (VP)
English-RAs	All	2	1	----	2	1
	It	41	16	34	73	5
	That	5	----	2	----	----
	Wh	14	----	3	----	2
Iranian-RAs	All	3	----	----	2	3
	It	12	3	10	71	----
	That	8	----	----	----	----
	Wh	8	----	1	----	1
Dissertations	All	4	----	1	----	4
	It	117	5	44	5	1
	That	2	----	6	----	----
	Wh	27	6	7	1	1
Textbooks	All	3	1	----	1	4
	It	178	12	115	67	----
	That	3	1	6	----	----
	Wh	62	----	12	6	7

Cleft sentences with PPs were further examined (Table 14). Prepositions employed by each group of writers were identified and categorized. The identification process was feasible because the number of prepositions is finite in the English language.

Table 14. Prepositions Used in Cleft Sentences

Prepositions	Frequency			
	English RAs	Iranian RAs	Dissertations	Textbooks
At	10	5	-----	8
Against	-----	-----	-----	1
During	2	6	-----	-----
For	11	2	-----	1
From	3	1	-----	-----
In	21	26	5	40

Into	-----	-----	-----	1
On	-----	1	-----	1
Through	14	25	-----	-----
To	2	1	1	8
Under	1	-----	-----	-----
Until	3	1	-----	13
Upon	1	-----	-----	-----
Via	-----	2	-----	-----
With	3	-----	-----	-----
Within	4	1	-----	-----

Here, we provide some extracts of for major types of cleft structures identified in our corpus. **All-clefts:** Textbook writers: Suddenly **all I** wanted to do **was** research the history of typewriters and typewriting. **It-clefts:** English RA writers: **It is** through the mediated struggle with contradiction **that** the activity **That-clefts:** Dissertation writers: **That is** why we write in all subject areas. **Wh-clefts:** Iranian RA writers: **What she** recommends **is** developing a "learning to learn competence".

Discussion

The analysis of cleft structures in RAs revealed that both English and Iranian writers employed cleft sentences in their RAs with no statistically significant difference. It-clefts were the most frequent cleft structure used by both groups of writers. The functions of it-clefts (such as focus bringing and contrast making) might be the reason behind the popularity of it-clefts compared to pseudo-clefts in academic writing. It can also be argued that it-clefts function similarly to existential *there (as in there is a significant difference).* One function that these two structures have in common is summarizing the given information (Biber et al., 1999; Jiang and Hyland, 2020). This feature of existential *there* allows the writer to shift the reader's attention to the points made earlier in the text (Hyland, 2019). Hence, just like the high frequencies of existential *there* (Jiang & Hyland, 2020), this similarity might explain the abundance   of it-clefts in academic genres. RA writers are limited in terms of space. This limitation means that writers need to make their statements as precise as possible. Statements made using it-clefts are to the point and precise (Biber et al., 1999). This feature of it-clefts might explain the higher frequencies observed in the present study.

Pseudo-clefts were also used by both groups of writers. The number of pseudo-clefts identified in both English and Iranian RAs was in the same range. Previous research (Biber et al., 1999; Collins, 2002; Patten, 2012) has shown pseudo-clefts are known to be more frequent in conversation, fiction, and informal writing. Hence, a lower number of pseudo-clefts was expected in the genre of RAs. The frequencies of cleft structures found in the RAs corpus are in line with the other studies that investigated cleft sentences (Biber et al., 1999; Collins, 2002; Hasselgård, 2014; Patten, 2012). In the dissertations and textbooks, the same pattern of cleft structure occurrence can be observed. It-clefts were the most frequent cleft sentences in both genres (with 172 occurrences in dissertations, and 372 in textbooks). As was the case with RAs, compared to it-clefts, the occurrence of pseudo-clefts was considerably lower in both genres. The number of all-clefts and that-clefts was almost the same in both genres. However, the usage of wh-cleft in textbooks was twice as their occurrence in dissertations (87 in textbooks vs. 42 in dissertations).

With 10,300,366 words (5,158,755 English and 5,141,611 Iranian), the total number of cleft structures identified in the RAs corpus is 322 (201 in the English section and 121 in the Iranian section). Even though the comparison between English and Iranian writers, regarding the frequency of employing cleft structures in their writing, was not statistically significant, the results could still be interpreted in a meaningful way. As the most frequently used structure by both English and Iranian writers in this study (see Table 4), it-clefts play an important role in academic writing. By allowing writers to make precise statements, making comparisons and contrasts, and shedding light on what they want readers to focus on, it-clefts perform many functions. Both groups of writers took advantage of these functions in writing their RAs. However, in terms of frequency, a difference could be observed in it-clefts used by English and Iranian writers. From the total of 264 it-clefts in the RAs corpus, 169 belong to the English section of the corpus, and 95 to the Iranian section (64% and 36% respectively). A possible reason for such a difference may be lower language proficiency among Iranian writers. Another reason for the observed difference might have to do with the functions of it-clefts themselves. The statements made using it-clefts are precise and to the point. However, many textbooks on academic writing recommend that writers distance themselves from statements by using a passive voice or hedges (Sword, 2012; Wallwork, 2011). A final reason for the findings regarding it-clefts in RAs might be the differences between English and Persian cleft structures. These differences could be easily observed in the following example: *Ali raft*. The English counterpart of this cleft structure would be: *It was Ali that left*. Despite functioning in the same manner (as focus and contrastive devices), they are very different in

their forms. This difference in form may or may not interfere with the writing style of Iranian writers (Bhela, 1999).

The number of pseudo-clefts in both writer groups is remarkably similar. The English-speaking writers employed 32 pseudo-clefts in their RAs while Iranian writers used 26 pseudo-clefts. Both writer groups used almost the same number of all-clefts and that-clefts. However, English writers used almost twice as many wh-clefts as their Iranian counterparts (19 wh-clefts in the English category and 10 in the Iranian category). Pseudo-clefts feature prominently in speech and informal writing. This behavior is consistent with their lower frequency found in RAs compared to it-clefts. With the limited space imposed on RAs by the journals, at least in applied linguistics, the lower number of pseudo-clefts used by the writers is justifiable. The word limit simply does not allow writers to add anything other than to-the-point statements in their RAs. However, if the small difference in the use of wh-clefts between English-speaking and Iranian writers is to be considered, it shows that English-speaking writers still prefer to add a more "personalized style" to their RA writing. Structures such as *what I mean is* or *what I want to emphasize is* allow writers to make conclusions and bring focus to the previous points they made in their arguments.

With a statistically significant number of cleft structures, the dissertation genre displayed the same pattern as the RA genre written by English-speaking speakers. It-clefts were much higher in frequency than their pseudo-cleft counterparts (172 it-clefts and 60 pseudo-clefts). The number of it-clefts dissertations is almost the same as those found in the RA genre written by English-speaking writers (169 it-clefts). It-clefts allow writers to make contrasts and precise statements, which render them suitable for academic writing. As Biber et al. (1999) noted, precise statements that can be made using cleft structures are suitable for academic writing. A similarity between RAs and PhD dissertations in terms of it-cleft occurrence is observed as well. This similarity may be due to the fact that many students turn and publish their dissertations into RAs as part of their graduation process (Lee, 2010). This practice is also observed in Iranian academic settings. Pseudo-clefts might be removed from the RAs in this process due to their informal nature. On the other hand, it-clefts might be kept for making precise and contrastive statements. Pseudo-cleft dissertations were considerably higher than those in RAs written by English-speaking writers. The observed difference might possibly be because of the extra space in dissertations compared to RAs. Not setting a word limit allows the writers to revisit their ideas and draw more conclusions and summaries.

Just like the dissertations, the same pattern of frequencies is observed in textbooks. With 372 it-clefts and 106 pseudo-clefts, the frequencies are considerably higher than those

in the other two genres. One reason for these differences might have to do with the fact that textbook authors usually have a greater mastery of their subjects. This mastery allows them to make more precise statements, draw more conclusions, and make more contrasts between their sentences. Another reason can be the nature of the textbook itself. Textbooks, especially those targeted toward BA and MA students, such as the ones included in this study, are made to get a point across (Swales, 1990). By using it-clefts, authors are able to draw the students' attention to the points they are making (Patten, 2012). Just as was the case with dissertations, textbooks also give the writers freedom to use an almost unlimited number of words. With more space to work with, writers might choose to use more cleft sentences in their textbooks. A final reason that might explain the larger number of it-clefts in textbooks compared to the other two genres might be the nature of textbooks. Also referred to as a hybrid or blurred genre (Swales, 1995), unlike RAs, textbook authors arrange the currently accepted knowledge into a coherent whole (Myers, 1992). Textbook writers try to make the material as clear as possible for their target audience. Using clefts might enable the textbook writers to regurgitate, summarize, and highlight their ideas in writing.

Conclusion and Implications

Comparisons between English-speaking and Iranian writers in terms of the frequency of clefting showed similar use of this construction. However, a difference, though not statistically significant, was observed between the two groups in the number of it-clefts and wh-clefts, implying that English writers use it-clefts and wh-clefts more liberally than Iranian writers. These findings indicate that Iranian RA writers are slightly falling behind their English native counterparts in using cleft structures. Both it-clefts and wh-clefts play important roles in directing readers' attention to the points made by writers. Therefore, Iranian writers should employ more direct, precise, and personal statements.

Although it-clefts were more frequently used in textbooks and dissertations, the role of wh-clefts should not be downgraded because they help writers add a "personalized touch" to their writing. In doing so, they reduce reliance on the use of passive voice, which according to some researchers (Belmont & Sharkey, 2011; Hall & Birkerts, 2007) is a desirable behavior in academic writing.

Although cleft constructions of all types tended to occur in all three academic genres in the present study in varying degrees, given the relatively large specialized corpus size of the present study (20,389,297), only about 4.5 cleft sentences recur in every 100,000 words. This low frequency of such grammatical structures appears to contract Biber et al.'s (1999)

observation that "clefting, unlike dislocation, is common both in conversation and in the written registers" (p. 936). Such a conclusion implies that it may not be worth the time, energy, and cost to explicitly instruct such constructions to student writers for academic purposes in the academia.

Notwithstanding the above reservations, the major pedagogical implication of the findings of the present study concerns novice, non-native English student writers who should familiarize themselves with complex grammatical structures such as clefts in order to make their writing look more professional. One of the safest ways to achieve this is to design consciousness-raising activities coupled with explicit instruction to help them notice such grammatical constructions in concordance lines for a better understanding of their forms and functions.

References

Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.laurenceanthony.net/software

Ball, C. N. (1991). *The historical development of the it-cleft* (AAI9125587) [Doctoral dissertation]. Retrieved from ProQuest Dissertations and Theses database (AAI9125587).

Ball, C. N. (1994). The origins of the informative-presupposition it-cleft. *Journal of Pragmatics*, *22*(6), 603-628.

Becher, T., & Trowler, P. (2001). *Academic tribes and territories* (2nd ed.). Open University Press.

Belmont, W., & Sharkey, M. (2011). *The easy writer: Formal writing for academic purposes*. Pearson Longman.

Bhela, B. (1999). Native language interference in learning a second language: Exploratory case studies of native language interference with target language usage. *International Education Journal, 1*(1), 22-31.

Biber, D. (1993). Representativeness in corpus design. *Journal of Literary and Linguistic Computing*, *8*(4), 243-257.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Grammar of spoken and written English*. Harlow: Longman.

Bruce, I. (2014). Expressing criticality in the literature review in research article introductions in applied linguistics and psychology. *English for Specific Purposes*, *36*(4), 85-96.

Brumback, R. A. (2009). Impact factor wars: Episode V: The empire strikes back. *Journal of Child Neurology*, *24*(1), 260–262.

Calude, A. S. (2007). *Demonstrative clefts in spoken English* [Master's thesis]. Retrieved from https://researchspace.auckland.ac.nz/handle/2292/2415

Calude, A. S. (2017). Sociolinguistic variation at the grammatical/discourse level. *International Journal of Corpus Linguistics*, *22*(3), 429-455. doi: 10.1075/ijcl.22.3.06cal

Chan, T. H. (2015). A corpus-based study of the expression of stance in dissertation acknowledgements. *Journal of English for Academic Purposes*, *20*(4), 176-191.

Collins, P. C. (2002). *Cleft and pseudo-cleft constructions in English*. Routledge.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century?. *TESOL Quarterly, 34*(3), 548–560. doi: 10.2307/3587743

Dong, Y. R. (1998). Non-native graduate students' thesis/dissertation writing in science: Self-reports by students and their advisors from two US institutions. *English for Specific Purposes*, *17*(4), 369-390.

Egbert, J. (2007). Quality analysis of journals in TESOL and applied linguistics. *TESOL Quarterly*, *41*(1), 157–171.

Goldberg, A. E. (2006). *Constructions at work*. Oxford University Press.

Hall, D., & Birkerts, S. (2007). *Writing well* (9th ed.). Longman.

Halliday, M. A. K. (1994). *An introduction to functional grammar*. Edward Arnold.

Hashemi, M., & Babaii, E. (2013). Mixed methods research: Toward new research designs in applied linguistics. *The Modern Language Journal*, *97*(4), 828-852. Doi: 10.1111/j.1540-4781.2013.12049.x.

Hasselgård, H. (2014). It-clefts in English l1 and l2 academic writing: The case of Norwegian learners. In K. Davidse, C. Gentens, L. Ghesquière, & L. Vandelanotte (Eds.), *Corpus interrogation and grammatical patterns* (pp. 295-320). John Benjamins.

Henderson, W., Dudley-Evans, T., & Backhouse, R. (1993). *Economics and language*. Routledge.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: issues and directions. *Journal of English for Academic Purposes*, *1*(1), 1-12. doi: 10.1016/s1475-1585(02)00002-4

Hyland, K. (2019). *Metadiscourse: Exploring interaction in writing*. Continuum.

Hyland, K. (2023). *Enter the dragon: China and global academic publishing*. Learned Publishing.

Jiang, F., & Hyland, K. (2020). There are significant differences…: The secret life of existential there in academic writing. *Lingua*, *233*(1), 1-17.

Jung, U. O. H. (2004). Paris in London revisited or the foreign language teacher's topmost journals. *System*, *32*(3), 357–361.

Larsen-Freeman, D. (1993). *Grammar dimensions: form, meaning, and use*. National Geographic Learning.

Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly, 34*(1), 175-181.

Lee, A. (2010). When the article is the dissertation. In C. Aitchison, B. Kamler, & A. Lee (Eds.), *Publishing pedagogies for the doctorate and beyond* (pp. 12-28). Routledge.

Leydesdorff, L., & Opthof, T. (2010). Scopus source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, *61*(1), 2365–2369.

Lim, J. M. H. (2014). Formulating research questions in experimental doctoral dissertations on Applied Linguistics. *English for Specific Purposes*, *35*(3), 66-88.

Lin, C. Y. (2017). I see absolutely nothing wrong with that in fact I think …: Functions of modifiers in shaping dynamic relationships in dissertation defenses. *Journal of English for Academic Purposes*, *28*(4), 14-24.

Lu, X., & Deng, J. (2019). With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. *Journal of English for Academic Purposes*, 39, 21-36.

Matikainen, T. (2024). Academic writing in English: Lessons from an EMI-program in Japan. *Journal of English for Academic Purposes*. Doi: 10.1016/j.jeap.2024.101358

Moghaddasi, S., & Graves, H. A. (2017). Since Hadwiger's conjection … is still open: Establishing a niche for research in discrete mathematics research article introductions. *English for Specific Purposes*, *45*(1), 69-85.

Myers, G. (1992). Textbooks and the sociology of scientific knowledge. *English for Specific Purposes*, *11*(1), 3-17.

O'Keeffe, A., & McCarthy, M. (2010). Historical perspective: What are corpora and how have they evolved?. In Michael McCarthy and Anne O'Keeffe, *The Routledge handbook of corpus linguistics* (pp. 31-41). Routledge.

Patten, A. (2012). *The English it-cleft: A constructional account and a diachronic investigation*. De Gruyter Mouton.

Prince, E. F. (1978). A comparison of wh-clefts and it-clefts in discourse. *Language*, *54*(4), 883-906. doi: 10.2307/413238

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M. (1995). The role of the textbook in EAP writing research. *English for Specific Purposes*, *14*(1), 3-18. doi: 10.1016/0889-4906(94)00028-c

Swales, J. M., & Post, J. (2018). Student use of imperatives in their academic writing: How research can be pedagogically applied. *Journal of English for Academic Purposes*, *31*(1), 1-97.

Sword, H. (2012). *Stylish academic writing*. Harvard University Press.

Thomson, A. J., & Martinet, A. V. (1986). *A practical English grammar* (4th ed.). Oxford University Press.

Wallwork, A. (2011). *English for Writing Research Papers* (Vol. 137). Boston, MA: Springer US.

Warchał, K. (2010). Moulding interpersonal relations through conditional clauses: Consensus-building strategies in written academic discourse. *Journal of English for Academic Purposes*, *9*(2), 140-150.

Weber, M., & Campbell, C. M. (2004). In other professional journals. *The Modern Language Journal*, *88*(1), 457–466.

Weiner, G. (2001). The academic journal: Has it a future? *Education Policy Analysis Archives*, *9*(1).

Zotzmann, K., & Sheldrake, R. (2021). Postgraduate students' beliefs about and confidence for academic writing in the field of applied linguistics. *Journal of Second Language Writing*, *52*, 100810.

Appendix: Settings Used in the AntConc Software

Cleft types	All-clefts	It-Clefts	That-Clefts	Wh-clefts
Search terms	All i|all he|all she|all they	it is|it was|it were	that is|that was|that were	what i|what he|what she|what they
Search Window Size	100	100	75	75
Context words	Are, is, that,	All, one, place,	He, she, I, we,	Are, is, that, was,

	was, were	reason, that, thing, time, way	researcher	were
Context words range	L = 0 R = 10	L = 0 R = 10	L = 0 R = 10	L = 5 R = 10

Note[1]. Search window size affects the number of concordance lines. The search window size was optimized through trial and error.

Note[2]. Uppercase or lowercase letters did not affect the search.