Research Article

Abdelhamid M. Ahmed*, Xiao Zhang, Lameya M. Rezk and Wajdi Zaghouani

Building an Annotated L1 Arabic/L2 English Bilingual Writer Corpus: The Qatari Corpus of Argumentative Writing (QCAW)

https://doi.org/10.1515/csh-2023-0012
Received July 19, 2023; accepted December 21, 2023; published online January 12, 2024

Abstract: The study presents the creation of the Qatari Corpus of Argumentative Writing (QCAW) as an annotated L1 Arabic and L2 English bilingual writer corpus. It comprises 200,000 tokens of argumentative writing by Qatari university students in L1 Arabic and L2 English. The corpus includes 195 essays written by 195 students, 159 females and 36 males. The students were native Arabic speakers proficient in English as a second language. The corpus is divided into Arabic and English sections, accompanied by part-of-speech annotated files in UTF-8 encoded text format. Metadata in CSV format contains information about the students (gender, major, first and second languages) and the essays (text serial numbers, word limits, genre, writing date, time spent, and location). The current study outlines the steps for collecting and analysing the corpus, including details on essay writers, topic selection, pre-analysis text modifications, proficiency level, gender, and major ratings. Statistical analyses were applied to examine the corpus. The QCAW offers a valuable bilingual data source authored by the same students in Arabic and English, with implications for further research.

***Corresponding author: Abdelhamid M. Ahmed**, Assistant Professor in English Language Education, Core Curriculum Program, Deanship of General Studies, Qatar University, Doha, Qatar, E-mail: aha202@qu.edu.qa. https://orcid.org/0000-0001-9667-8630
Xiao Zhang, Postdoctoral Researcher, Core Curriculum Program, Deanship of General Studies, Qatar University, Doha, Qatar
Lameya M. Rezk, Postdoctoral Researcher, College of Humanities & Social Sciences, Hamad Bin Khalifa University, Doha, Qatar
Wajdi Zaghouani, Associate Professor of Digital Humanities, College of Humanities & Social Sciences, Hamad Bin Khalifa University, Doha, Qatar

Keywords: argumentative writing; corpus building; L1 Arabic; L2 English; Qatari students – Qatari Corpus

1 Introduction

Learner corpora are authentic language data produced by individuals learning their first or second language (Granger, Gilquin, and Fanny 2015; Gilquin, Granger, and Paquot 2007). Granger (2003) views learner corpora as a novel resource for specialists in Second Language Acquisition (SLA) and Foreign Language Teaching (FLT). Additionally, learner corpus research is situated at the intersection of four significant disciplines: corpus linguistics, linguistic theory, SLA, and FLT, as highlighted by Granger (2009).

Previous research highlighted the significance of and the need to create learner corpora. Firstly, knowledge derived from learner corpora can have significant pedagogical implications by prioritising specific vocabulary classes, including multiword clusters that learners underuse (Shirato and Stapleton 2007). Secondly, Dashtestani and Stojkovic (2016) found that learner corpora can enhance students' academic vocabulary, word combination learning, and communicative abilities. Thirdly, learner corpora are essential for identifying and quantifying common error types, prioritising the development of error-specific algorithms, providing training data for machine-learned approaches, and evaluating error detection and correction systems, as argued by Gamon et al. (2013). Moreover, learner corpora are crucial in expanding the Interlanguage Pragmatics (ILP) limited research agenda, as Callies (2013) noted. Student feedback also suggests that learners find using corpora beneficial even with their limited English proficiency, as Okamoto (2010) reported.

Furthermore, Gilquin, Granger, and Paquot (2007) highlight that learner corpora are useful in English for Academic Purposes (EAP) pedagogy since they expose issues non-native learners face while writing academic essays. Additionally, learner corpora offer learners more exposure to authentic examples, making them valuable resources for pedagogic purposes, from syllabus design to materials development, as emphasised by Kayaoglu (2013). The current study discusses how the Qatari Corpus of Argumentative Writing (QCAW) was built as an annotated L1 Arabic and L2 English bilingual writer corpus.

1.1 Research Aims

The aims of the current research are threefold. The first aim of this study is to develop a bilingual corpus, referred to as the Qatari Corpus of Argumentative Writing (QCAW), comprising 200,000 tokens of argumentative writing produced by

Qatari university students in both L1 Arabic and L2 English. This aim focuses on the three critical stages of corpus construction: corpus design, development, and annotation. The second aim of the present study is to highlight how the QCAW enhances our understanding of texts written by the same bilingual writers in both L1 Arabic and L2 English, specifically metadiscourse markers. The last aim of the current study is to shed light on some specific applications derived from the Qatari Corpus of Argumentative Writing (QCAW) to enhance our understanding of bilingual writing in both L1 Arabic and L2 English contexts, with specific reference to form variability, learner errors, and writing assessment.

1.2 Research Significance

The current study is of considerable scientific significance for multiple reasons. First, building the Qatari Corpus of Argumentative Writing (QCAW) addresses a discernible gap in existing research by creating the first bilingual corpus in L1 Arabic and L2 English that encapsulates the nuanced dynamics of argumentative writing in both languages. This innovative approach provides a distinctive opportunity to explore the intricate interplay of linguistic features – ranging from vocabulary use to metadiscourse markers, syntactic structures, and part-of-speech distribution, writing errors – within the same set of writers across two languages. Secondly, the study contributes to advancing corpus linguistics, a pivotal field for understanding language patterns and structures. The meticulous annotation of the corpus, incorporating Part-Of-Speech (POS) annotations, enhances the scientific rigour of linguistic analyses, deepening our comprehension of specific linguistic features in argumentative writing and offering a valuable resource for developing and refining theories applicable to bilingual contexts.

Moreover, the demographic analysis, including considerations of gender and major in relation to writing proficiency, aligns with the broader goals of educational research, contributing to a more comprehensive understanding of factors influencing writing proficiency in bilingual university students. Additionally, the detailed procedural aspects of corpus collection and analysis outlined in the study contribute methodologically to the field, providing valuable insights into best practices for corpus creation and analysis and enhancing the robustness and replicability of future research endeavours. Lastly, the study's exploration of the implications of the QCAW as a bilingual data source is of paramount importance. By emphasising its potential contributions to language acquisition, argumentative writing, and bilingual studies in both L1 Arabic and L2 English contexts, the research extends the practical applications of its findings, fostering advancements in pedagogy and linguistic research. In summary, the scientific justification for this study resides in its distinctive contribution to the fields of corpus linguistics, bilingual education,

and educational research, presenting a valuable resource for scholars, educators, and researchers interested in the intricacies of argumentative writing proficiency in a bilingual setting.

2 Literature Review

The literature review section discusses learner corpora, exploring their various types, including written and spoken learner corpora, learner-compared corpora, L2 English learner corpora, and L1 learner corpora. These corpora differ based on collection time, scope, targeted language (L2), learners' mother tongue (L1), medium, and text type. The most common text types represented in learner corpora are argumentative texts for writing and informal interviews for speaking. The section also examines the available Arabic–English bilingual corpora and highlights the features of L1 Arabic and L2 English writing. This review aims to provide valuable insights into language learning processes and interlanguage development.

2.1 Types of Learner Corpora

There are six different types of learner corpora, each with unique characteristics and uses. Firstly, the written learner corpora consist of written texts produced by language learners, such as essays, journals, and emails (Coxhead 2000; Gilquin and Granger 2015). These corpora are useful for studying language learners' errors, error patterns, and language development over time. Secondly, the spoken learner corpora consist of spoken language produced by language learners, such as oral interviews, dialogues, and conversations (Caines, McCarthy, and O'Keeffe 2016; Yoon 2020). These corpora study language learners' pronunciation, fluency, and spoken discourse strategies. Thirdly, the learner-compared corpora consist of written or spoken texts produced by language learners and native speakers, allowing for a direct comparison of language use between the two groups. These corpora are useful for identifying the specific areas in which language learners struggle and for identifying patterns of language use unique to language learners (Gilquin, Granger, and Paquot 2007).

In addition, learner corpora differ in multiple dimensions, including the time of collection, the scope of the collection, the targeted language (L2), the learner's mother tongue (L1), the medium, and the text type (Granger 2011). In reference to the time of collection, there are two types of learner corpora: cross-sectional learner corpora and longitudinal learner corpora. The former consists of instances of learner writing or speech collected from various categories of learners at a particular moment. In contrast, the latter monitors the progress of identical learners over

a specific time frame. In relation to the scope of the collection, two types of learner corpora are identified: global and local. Global learner corpora are large data collections from diverse learners that inform SLA theory and teaching tools. On the other hand, local learner corpora are smaller collections gathered by teachers in their routine teaching practices, used as the foundation for classroom materials.

Another way to categorise learner corpora is based on the language they focus on, such as L2 English learner corpora and L1 learner corpora. In terms of the medium, there is a written learner corpus, which refers to corpora of learner writing. In contrast, a spoken learner corpus may refer to transcriptions of oral production data. Finally, based on the text types, the two most commonly represented text types in learner corpora are argumentative texts for writing and informal interviews for speaking.

2.2 Differences Between Arabic and English

Salloum et al. (2023) outlined eight notable distinctions between Arabic and English. The first contrast lies in character connections: English utilises diagonal strokes to link characters, while Arabic connects the baseline with horizontal strokes. Second, English character versions exhibit limited shape variations. In contrast, Arabic characters display significant variability, showcasing up to four distinct shapes based on word position. Capitalisation is a feature exclusive to English. The direction of writing diverges, with English following a left-to-right pattern and Arabic adopting a right-to-left orientation. Additional differences include gender differentiation in Arabic verb and sentence structure, plural forms (singular and plural in English versus singular, dual, and plural in Arabic), adjective placement (before the noun in English and after in Arabic), and segmentation methods in handwriting. Moreover, the alphabet size differs (26 letters in English and 28 in Arabic), and sentence types vary (verbal in English and nominal and verbal in Arabic). Lastly, the total number of speakers is substantial, with English at 1.348 billion and Arabic at 274 million.

2.3 Arabic–English Bilingual Corpora

The Zayed Arabic–English Bilingual Undergraduate Corpus (ZAEBUC) corpus is the only Arabic–English bilingual corpus available online. The ZAEBUC corpus was developed by Habash and Palfreyman (2023). It comprises bilingual writing samples from the same writers on different occasions, matching comparable texts in different languages. Specifically, it currently contains short essays from several hundred Freshman students, predominantly Emirati. The corpus includes 388 English essays (88,000 words) and 214 Arabic essays (33,000 words).

The Qatari Corpus of Argumentative Writing (QCAW), under investigation, was published in the Linguistic Data Consortium by Ahmed et al. (2022). It comprises writing samples in L1 Arabic and L2 English written by the same Qatari students on two different Argumentative topics. It shows the same Qatari university students' argumentative writing in L1 Arabic and L2 English. It includes 195 essays in L1 Arabic (97,248 tokens) and 195 in L2 English (98,379 tokens). The next section sheds light on the features of L1 Arabic writing and L2 English writing.

2.4 Features of L1 Arabic Writing

Arabic written language is characterised by distinctive features that set it apart from other languages. Kaye (2017) identified Arabic as a Semitic language spoken by over 200 million people as a mother tongue. Arabic speakers primarily live in Southwest Iran, Iraq, Syria, the Arabian Peninsula, the Maghreb region of North Africa, Egypt, and Mauritania (Al-Khatib 2000). The Arab world is considered a diglossic speech community, where the language has two forms: colloquial Arabic, which exists as the vernacular varieties of the major Arabic-speaking nations, and classical Arabic, the language of the Quran, which provides a common, standard written form for all the vernacular variants and a shared medium for state affairs, religion, and education across the Arab world (Al-Khatib 1988, 1995).

The Arabic script comprises a set of 28 letters, each of which can take different forms depending on its position in the word (Khorsheed 2002). Additionally, Arabic script includes diacritical marks that indicate vowel sounds not represented in the script (Hamed and Zesch 2017). In addition, Arabic script uses ligatures, which are combinations of two or more letters written as a single unit (Naz et al. 2016). Arabic grammar includes two genders (feminine and masculine), three numbers (singular, dual, and plural), and three grammatical cases (nominative, genitive, and accusative) (Chen and Gey 2002).

Arabic written language is marked by its use of the definite article, represented by the prefix "al-" (Al-Jarf 2022). This noun prefix indicates that it is definite and changes the form of the noun depending on its grammatical case (Chen and Gey 2002). Besides, Arabic has a complex grammatical structure, with a system of nouns, verbs, and other parts of speech that are inflected to indicate tense, mood, and other grammatical features (Sawalha and Atwell 2013).

Arabic written language has an inflexion system, known as declensions, which indicate the grammatical function of nouns and adjectives (Saiegh-Haddad and Henkin-Roitfarb 2014). This system depends on the use of patterns of consonants and vowels, which change depending on the grammatical case and the number of words (Abu-Rabia and Awwad 2004). The inflexion system is clear in Arabic nouns, which have three different cases (nominative, genitive, and accusative) and three

different numbers (singular, dual, and plural) (Chen and Gey 2002). Another characteristic of the inflexion system is that Arabic verbs have a complex conjugation system based on the person, gender, and number of subjects (Kusters 2003).

In addition, the written Arabic language also includes a set of grammatical particles known as particles of negation, which are used to indicate negation and other grammatical functions (Al-Momani 2011). Furthermore, the Arabic written language has a rich system of idiomatic expressions, proverbs, and colloquial expressions, which convey meaning and emphasise certain ideas (Alqahtni 2014).

To summarise, Arabic has a distinctive script (the Arabic alphabet). It uses a complex system of declensions and particles of negation. It also has a rich tradition of idiomatic expressions, proverbs, and colloquial expressions, making it a unique and complex language.

2.5 Features of L2 English Writing

L2 English writing is characterised by some features different from native speakers. Grammatical, lexical, syntactical and orthographic errors are prevalent in L2 English learners' writing (Olsen 1999). For example, Arab students often struggle with L2 English grammar, vocabulary, organisation and coherence in their English writing (Khuwaileh and Shoumali 2000). These errors are often caused by the influence of the learners' first language (L1) on their second language (L2) (Crompton 2011). These errors may also be attributed to students' problems with the cultural and linguistic differences between their native language and English (Al-Jarf 2013).

Overgeneralization is another feature of L2 English learners in writing. It occurs when learners apply the rules of their L1 to the L2 (Mourssi 2013). Overgeneralization is particularly common in L2 English learners using irregular verb forms and verb tenses (Kirmizi and Karci 2017). Additionally, learners may overgeneralise grammatical structures from their L1, such as articles or word order (Hertel 2003).

Many English learners have a limited vocabulary, sometimes resulting in repetitive words and phrases (Ahmed 2010a). Learners' limited vocabulary repertoire may lead to problems with word order and collocation, showing an insufficient command of more complex vocabulary that enables them to express their ideas precisely (Phoocharoensil 2013). Additionally, L2 English learners have problems with coherence, cohesion, lexis, grammar and mechanics (Ahmed 2010b), making it difficult for readers to understand the intended meaning. These problems are attributed to socio-cultural issues (Ahmed and Myhill 2016). Moreover, English learners may also have difficulties using cohesive devices such as referencing, substitution, and ellipsis, which are crucial for text coherence (Ahmed 2010b).

2.6 Research Questions

1. How was the Qatari Corpus of Argumentative Writing (QCAW) built in terms of corpus design, development and annotation?
2. How does the Qatari Corpus of Argumentative Writing (QCAW) enhance our understanding of texts written by the same bilingual writers in both L1 Arabic and L2 English, specifically metadiscourse markers?
3. What specific applications can be derived from the Qatari Corpus of Argumentative Writing (QCAW) in enhancing our understanding of bilingual writing in both L1 Arabic and L2 English contexts, with specific reference to form variability, learner errors, and writing assessment?

3 Methodology

This section highlights how the Qatari Corpus of Argumentative Writing (QCAW) was built, taking into consideration the corpus design, corpus development and corpus annotation stages.

3.1 Corpus Design

3.1.1 Scope and Objectives

The Qatari Corpus of Argumentative Writing (QCAW) design involved meticulous planning to ensure representation and relevance. Comprising 390 essays in L1 Arabic and L2 English, the corpus aimed to create a comprehensive dataset that captures Qatari university students' diverse argumentative writing styles. This involved defining inclusion criteria, selecting appropriate essay topics, and ensuring the sample's representativeness.

3.1.2 Topic Selection and Questionnaires

Ten argumentative prompts were derived from TOEFL essay writing prompts to select suitable topics. Questionnaires were designed for both students and instructors to assess their preferences. The top two topics were chosen through collaborative responses from six instructors and 34 students, forming the basis of the final prompts. This process ensured that the topics were relevant and engaging for the participants.

3.1.3 Task Allocation and Pilot Study

Students were divided into groups A and B to mitigate task and topic effects. Group A tackled Topic 1 in Arabic and 2 in English, while Group B did the reverse. This allocation ensured a balanced representation in the corpus. Before the full-scale data collection, a pilot study involving six students assessed the appropriateness of the conditions, including task time, word count, and clarity of instructions. Adjustments were made based on the feedback received.

3.1.4 Target Beneficiaries of the Corpus

The primary focus here is identifying the intended users of the Qatari Corpus of Argumentative Writing (QCAW). The section emphasises educators, researchers, and corpus linguists as the primary audience, indicating the corpus role in supporting pedagogical development, language acquisition studies, and broader research inquiries in bilingual education. This aligns with the design phase, where the purpose and potential impact of the corpus are delineated to guide its development.

3.2 Corpus Development

3.2.1 Data Collection and Demographics

The data collection phase involved students aged 18 to 22 enrolled in a compulsory First-Year Seminar course. The bilingual students, with Arabic as their L1 and English as their L2, were instructed to write argumentative essays in both languages. A focus on gender balance was maintained, aligning with the university's female-to-male student ratio of 3 to 1. Considering the diversity of majors and years of study, the sample's representativeness was crucial.

3.2.2 Topic Implementation and Feedback

After selecting topics and dividing students into groups, the actual implementation involved students writing argumentative essays based on their knowledge and experiences. The process was iteratively refined through feedback sessions and a pilot study to ensure the conditions were suitable for all participants. This phase aimed to collect high-quality, representative data for subsequent analysis.

3.2.3 Benchmarking Procedures

Benchmarking procedures were implemented to maintain the reliability of the collected data. Raters underwent training sessions, norming sessions, and independent rating sessions. These steps were crucial in establishing a shared understanding of the assessment criteria, ensuring consistency in the evaluation of writing quality and voice salience.

3.2.4 Standardising Argumentative Writing

Standardising argumentative writing within the QCAW establishes a benchmark for evaluating and understanding bilingual university students' proficiency levels. It is a crucial step in the development phase as it outlines the methodologies employed to ensure consistency and reliability in the collected data. The standardisation process creates a reference point for comparative analyses, indicating strengths and areas for improvement in bilingual writing skills.

3.2.5 Representativeness and Acknowledgment of Bias

In this phase, attention is given to ensuring the collected data is diverse and reflective of the population under study. The discussion on gender distribution and the acknowledgement of the 3-1 female-to-male student ratio at the university in Qatar signifies the conscientious effort made during data collection to maintain a semblance of balance. Furthermore, recognising observed errors within the corpus and their potential impact on linguistic challenges faced by bilingual writers aligns with the ongoing scrutiny required for developing a robust and insightful corpus. This section contributes to the transparency and validity of the corpus development process.

3.2.6 Corpus Building Challenges

The following lines highlight the challenges of building a unique Arabic/English argumentative writing corpus. We encountered some challenges while building the Qatari Corpus of Argumentative Writing (QCAW). Firstly, selecting the writing topics for students to develop an argument easily was challenging. We consulted TOEFL Writing topics and selected ten topics. We surveyed students and instructors about their preferred topics in a questionnaire to select two topics. Secondly, it was a real challenge to motivate Qatari university students to write two essays, one in Arabic and another in English, for a non-graded assignment. A few students volunteered to do so. However, we contacted some instructors who motivated their

students to complete the task for us. Thirdly, students wrote their essays in two classes under controlled conditions without access to external resources. This was a challenge as some students wanted to access the internet to look for supporting ideas for their arguments. Fourthly, some students preferred to have their essays hand-written and not typed. This challenge was time-consuming as we had to type all these essays with all their mistakes in Arabic and English. Additionally, obtaining the ethical approval certificate from the concerned university took over three months. Getting students to read and voluntarily sign their consent forms was time-consuming. Another challenge was male versus female representation in the corpus sample. We collected essays from 154 female students versus 41 male students due to the 3-1 female-to-male student ratio at the concerned university in Qatar. Furthermore, reaching some exclusion criteria for the corpus texts was challenging. Finally, the research team was challenged with rating student essays for writing proficiency and voice in L1 Arabic and L2 English.

3.3 Corpus Annotation

3.3.1 Linguistic Annotation Tools

The QCAW underwent thorough annotation processes to enhance its linguistic value. Part-of-speech (POS) annotation was applied using internationally recognised tools, such as TreeTagger for English and Farasa for Arabic texts. These tools, known for their accuracy, ensured consistent and reliable annotations.

3.3.2 Metadata Inclusion

To enrich the corpus, metadata in CSV format was included, containing detailed information about the students (gender, major, first and second languages) and the essays (text serial numbers, word limits, genre, writing date, time spent, and location). This metadata serves as valuable contextual information for subsequent analyses.

3.3.3 Cleaning and Standardisation

Text cleaning involved applying exclusion criteria. In the final stage of corpus preparation, several texts were excluded based on four specific criteria. Essays with fewer than 250 words were excluded to prevent the inclusion of very short argumentative essays and potential inflation of values. Additionally, essays not responding to the writing prompt, those written only in Arabic, and those taking more than

50 min per essay were excluded. Hand-written texts were manually typed by the team members, resulting in a balanced set of 195 English and 195 Arabic essays.

The Arabic and English texts underwent further annotation and amendments before analysis. Headings, titles, and repeated task instructions or writing prompts were removed. Spelling corrections were manually applied to capture intended metadiscourse features, standardising to American English. However, no corrections were made for grammatical accuracy or turn of phrase. The raw and amended texts are accessible on the corpus website (https://catalog.ldc.upenn.edu/LDC2022T04).

The Qatari Corpus of Argumentative Writing (QCAW) employed Part-of-Speech (POS) annotation in UTF-8 encoded text format. Metadata in CSV format provided detailed information about students and essays, including gender, major, first and second languages, text serial numbers, word limits, genre, writing date, time spent, and location. English texts were annotated using TreeTagger, and Arabic texts using Farasa, both internationally recognised tools for accuracy. Both raw and annotated texts are made available for all users. Table 1 summarises the English-Arabic corpus make-up, detailing text counts, average essay lengths, standard deviations, essay length ranges, and corpus tokens for both languages.

3.4 Rating Essays for Writing Quality and Voice

In the corpus annotation stage, both English and Arabic texts underwent separate rating processes to assess writing quality and dimensions of voice. The essays were initially graded for writing quality using a five-category analytical rubric, emphasising students' stances over structural elements. A benchmarking procedure involving four raters ensured reliability, with alignment checks and discussions to maintain consistency. For voice assessment, a holistic voice rubric was employed, scored by four native Arabic-speaking raters, including both Egyptian and Tunisian perspectives. Benchmarking procedures for voice included norming sessions, consensus-building discussions, and reliability measures to ensure consistent evaluations. Raters focused solely on assessing voice salience, and the general instructions emphasised reliability through breaks, rubric reviews, and double-rating each writing sample. These rigorous procedures contribute to the validity

Table 1: English–Arabic corpus make-up.

Corpus	Texts	Average essay length	SD	Essay length range	Corpus tokens
Arabic	195	498.71	84.56	251–808	97,248
English	195	504.51	94.87	263–1158	98,379

of the assessments and enhance the overall quality of the Qatari Corpus of Argumentative Writing (QCAW).

3.5 Standardising Argumentative Writing

We standardised argumentative writing within the QCAW to establish a benchmark or reference point for evaluating and understanding bilingual university students' proficiency levels in argumentative writing in L1 Arabic and L2 English. It enables comparative analyses to highlight areas of strength and improvement in bilingual writing skills.

3.6 Research Limitations

While the current study aims to create a robust bilingual corpus for analysing argumentative writing, some limitations should be acknowledged. Firstly, the students participating may exhibit a wide range of proficiencies in both L1 Arabic and L2 English. This variability in language capabilities could impact the quality and sophistication of argumentative writing samples collected, introducing potential variation. Secondly, some students might be more accustomed to the argumentative writing task assigned based on prior coursework or experience, while it may be new for others. This could influence the structure and development of their written arguments. Thirdly, L1 Arabic language structures might interfere with or influence students' L2 English writing and vice versa. Such cross-linguistic influences could contribute to variations in writing style, rhetorical preferences, or organisational approaches. Finally, the in-class timed writing situation, while well-controlled, limits student access to online resources, references, or assistance compared to a take-home writing assignment. The absence of such support conceivably has some impact on the writing process.

While the Qatari Corpus of Argumentative Writing incorporates rigorous design to ensure a robust dataset, these limitations acknowledge potential sources of variability stemming from the student writers, which could impact generalizability. Further studies controlling for some of these factors would prove valuable.

3.7 Ethical Considerations

Several measures were implemented to ensure this study was conducted ethically and safeguard participants' rights. Approval was obtained from the university Institutional Review Board (IRB) overseeing studies involving human subjects before any participant recruitment or data collection.

Once approved, student participants were given information sheets outlining the study's purposes and procedures in accessible language. Informed written

consent was collected from all participants, specifying that their involvement was voluntary with no impact for declining. Steps were taken to allow anonymous participation, with all identifying information removed from essays during corpus compilation, and participants were assigned ID numbers. Students could withdraw their consent at any time without repercussion.

Collected written data was securely stored in encrypted files, accessible only to the research team members. Future users of the corpus for research purposes are appropriately bound by end-user agreements prohibiting attempts to identify participants. Any excerpts selected for publication are subject to stringent anonymisation practices.

This comprehensive approach ensured that participant anonymity, autonomy and confidentiality were preserved and maintained throughout the project in alignment with prevailing ethical standards for educational research. The procedures received full board approval, signifying adherence to established ethical guidelines for working with student subjects.

4 Research Findings

This section reports on the research findings in response to the three research questions of the current study.

4.1 QCAW Building

The answer to the first research question was answered in depth in the methodology section of this research paper, where the Qatari Corpus of Argumentative Writing (QCAW) went through the stages of corpus design, development and annotation.

4.2 QCAW and Metadiscourse Markers

In an attempt to answer the second research question, following Hyland's model of metadiscourse (2010), we have created Tables 2 and 3 to highlight how QCAW can be used to enhance our understanding of interactive and interactional metadiscourse categories with their specific functions, sub-categories and examples in both L1 Arabic corpus and L2 English corpus of Qatari university students.

How does the Qatari Corpus of Argumentative Writing (QCAW) enhance our understanding of texts written by the same bilingual writers in both L1 Arabic and L2 English, specifically metadiscourse markers?

Table 2 outlines the interactional category of metadiscourse, showing their functions, and provides examples in L1 Arabic and L2 English corpora extracted from the Qatari Corpus of Argumentative Writing (QCAW). The examples extracted

Table 2: Analysis of interactive metadiscourse markers in QCAW.

Category	Function	Examples of the Arabic Corpus	Examples of the English Corpus
Interactive	**Help to guide the reader through the text.**		
Transitions	Express semantic relations between main clauses: Addition, Compare/Contrast, Consequence	روابط الإضافة اعتماد المناهج الدراسية الجديدة وطرق الدراسة الحديثة على البريد الالكتروني وأيضا على استعمال الهواتف الذكية في بعض من الأمور والواجبات الدراسية. (131) روابط التضاد بالرغم من أن التكنولوجيا جعلت الأسار أكثر راحة، إلا أن الأضرار الناتجة عن ذلك كبيرة. (57) روابط النتيجة ومن ثم ينعكس على التنمية في القدرات الذهنية والفكرية للطلاب. (A177)	**Addition** Also, use programs and social networking to exchange of knowledge help to learn effectively. (110) **Compare/Contrast** However, sometimes students, and especially teenagers should have a limited access for technology and be allowed to it only when it benefits their studies and learning process. (117) **Consequence** Therefore, a lot of students use social networking programs for example, WhatsApp and Blackboard to help them. (122)
Frame Markers	Refer to discourse acts, sequences, or text stages.	ثانيا ، استخدام الهاتف الجوال والبريد الإلكتروني أصبح بديلا إيجابيًا عن الزيارات العائلية. (112)	Firstly, Technology is a great device that can assist and remodel education in many ways. (126)
Endophoric Markers	Refer to information in other parts of the text.	التواصل عن طريق التكنولوجيا أدى إلى قلة التواصل الاجتماعي بين الناس ويدعمون رأيهم بالحجج الآتية. (37A)	(146) To sum up, it clear that technology can benefits both students and teachers to study and to help them in many ways as mentioned above.
Evidentials	Refer to source information from other texts.	عن أنس بن مالك رضي الله عنه، قال سمعت رسول الله صلى الله عليه وسلم يقول "من سره أن يبسط له في رزقه أو ينسأ له في أثره فليصل رحمه". (173A)	None
Code Glosses	Help readers grasp the meaning of ideational materials.	على سبيل المثال، يوجد بين الناس أفراد انطوائيون لا يتحملون الاجتماعات المتكررة. (112A)	(158) For example, calls now are free through applications such as, Viber or Whatsapp calls.

N.B. All provided L1 Arabic and L2 English examples are verbatim from students' texts in the QCAW without corrections.

from QCAW illustrate how these metadiscourse markers operate in the respective languages. The interactive category includes the following metadiscourse sub-categories: transitions, frame markers, endophoric markers, evidentials, and code glosses.

Table 3 outlines the interactional category of metadiscourse, showing their functions, and provides examples in L1 Arabic and L2 English corpora extracted from the Qatari Corpus of Argumentative Writing (QCAW). The examples extracted from QCAW illustrate how these metadiscourse markers operate in the respective languages. The interactional category includes the following metadiscourse sub-categories: hedges, boosters, attitude markers, engagement markers, and self-mentions.

Table 3: Analysis of interactional metadiscourse markers in QCAW.

Category	Function	Examples of the Arabic Corpus	Examples of the English Corpus
Interactional	**Involve the reader in the tex**		
Hedges	Withhold the writer's full commitment to the proposition.	ومن الممكن أيـ ضا استخدام محركا ت البـ حث للمساعدة في البـ حوث العلمية. (113A)	(175) This may lead for that student to fail in his or her homework, or to get bad marks in an exam.
Boosters	Emphasise force or writer's certainty in the proposition.	وبـ التالي فإن الطالب يتمكن من تجميع المعلومات بـ صورة واضحة و سهلة. (113A)	(192) No one denies that books and libraries are interesting, and also a good place to spend time in.
Attitude Markers	Express writer's attitude to proposition	أوافق مع العبـا رة السابـقة بتهيـمن الهواتف الجوالة الحديثة بـشكل كبير على حياتنا اليـومية. (118A)	(201) Therefore, I agree that technology helps students to learn more information and to learn them fast.
Engagement Markers	Explicitly refer or build a relationship with the reader (e.g. Personal pronouns, questions...etc.).	فهل ساهمت التكنولوجيـ ا حقا في عمليـ ة الحصول على المعلومات بيسر وسهولة أم لا بـ ا ترى ؟ (4A)	(243) Could Technology Help the students to gain more information?
Self-mentɒns	Explicitly refer to authors.	أنا أتفق بـ شدة مع استخدام الهواتف ورسائل البريد الإلكتروني التي جعلت التواصل أقل شخصية. (90A)	(217) And I will support my opinion with reasons and examples in the following argument.

N.B. All provided L1 Arabic and L2 English examples are verbatim from students' texts in the QCAW without corrections.

4.3 QCAW Applications

This section attempts to answer the last research question stated below. What specific applications can be derived from the Qatari Corpus of Argumentative Writing (QCAW) in enhancing our understanding of bilingual writing in both L1 Arabic and L2 English contexts, with specific reference to form variability, learner errors, and writing assessment?

4.3.1 Applications of the Qatari Corpus of Argumentative Writing (QCAW)

Interpretation of learner corpus data differs from native language in many respects, among which the most important are form variability, learner errors, and writing assessment. In the case of learner English studies, form variability is usually

investigated in terms of variations of linguistic forms or language-specific variations. The term learner errors in this paper refers to using a word or an expression to denote a different meaning from native speakers, which could cause misinterpretation, ambiguity, or illogical statements. Writing assessment is inseparable from interpreting learner corpus data. It is especially applicable for teachers and researchers because written texts are a stable source for investigating the longitudinal progress of L1-specific learners and improving teaching strategies. To contribute to the aspects above, we will demonstrate how to use QCAW as a source of a learner corpus.

4.3.1.1 Form Variability in Learner English Corpus

Form variability is important in learner corpus research for two reasons. First, variations become unneglectable when the fact is "that the number of non-native speakers far outnumbers that of native speakers" (Granger, Gilquin, and Fanny 2015, p. 1). This results in an enormous number of language users, leading to higher variability in language forms. Second, to determine and describe the proficiency levels of learners, it is crucial to be aware of the differences or changes in language forms developed by the learners (Ädel 2015; Gilquin and Granger 2015; Gries and Wulff 2020; Hendriks 2005; Jarvis 2000; Mollin 2006; Paquot and Fairon 2006; Pendar and Chapelle 2008; Regan 2013; Vyatkina 2013; Wulff and Gries 2021).

One of the most common form variabilities in learner English is L1 specific variations, which are also our focus in this section. L1 specific variations, as the name suggests, refer to unique patterns found in L2 production of learners from specific L1 language backgrounds. They are not seen in the language use of native speakers. L1 specific variations vary from different language groups, i.e., variations used by one language group may rarely be seen in the language production of other language groups.

The QCAW is used here to exemplify L1 specific variations used by L1 Arabic speakers in their L2 English argumentative writing. Unique uses that occur more than twice are considered to be variations. Underlined parts in examples (1) to (3) illustrate the phenomenon of L1 specific variations.

(1) On the first hand, many people agree that emails and telephones are the best ways of communication in our life. [48B]

(2) On the first hand, the first team believes that nowadays technology is the only way to get any kind of information because it is easier, and not only that it is easier but also it is fast and quick with that students are more likely to use the technology whether it is on their mobile or tablets or even computers. [199B]

(3) On the first hand, it is believed that students need technology to further expand their knowledge because of the fact that it is very easy to access. [52B]

As the above examples show, “on the first hand” seemed inappropriate. However, they were not incorrect because they were coherent at the discourse level and did not interfere with readers’ understanding. Examples of another transition marker, “on the one hand”, are shown in Figure 1 below to supplement this case.

We further compared instances found in QCAW with argumentative writing by American university students provided by The Louvain Corpus of Native English Essays (LOCNESS).[1] Unsurprisingly, we found that “on the first hand” was not used by native speakers.

Another example of L1 specific variations found by our comparison between QCAW and LOCNESS-USARG is shown in examples (4)–(6).

(4) Some people say that technology affected education positively. On the other side, some people disagree to this and state that technology affected education in a bad way. [17B]

(5) On the other side, some people tell that the application of means of technology to education waste students time as some of them get to social media websites while being in classes instead of searching for information. [17B]

(6) On the other side, the people who oppose preferring studying on the technology support their point of view by many supports. [152B]

We found that L2 learners and native speakers used “on the other hand” to introduce a contrasting point of view in their argumentative writing. But, its equivalent, “on the other side”, was used only by L2 learners. This was also an example of L1 specific variation similar to examples (1)–(3). Figure 2 shows examples of “on the other hand” used by L2 learners.

L1 specific variations not only contribute to the interpretation of learner corpus data, they also render visible the differences between natives and non-natives. Furthermore, L1 specific variations are text-internal measures; they can be necessary to analyse the lexical diversity of L2 texts with quantitative data.

	File	Left Context	Hit	Right Context
1	40B.txt	show my opinion in the last part of the argument.	On the one hand,	there is a big number of
2	164B.txt	personal in ways as well as public in other ways.	On the one hand,	telephones and emails have made conversations
3	179B.txt	some of these views and express my view as well.	On the one hand,	the supporters say that technology always

Figure 1: Examples of “on the one hand”.

1 The Louvain Corpus of Native English Essays: https://www.learnercorpusassociation.org/resources/tools/locness-corpus/.

	File	Left Context	Hit	Right Context
1	12B.txt	meet them in person. It makes the communication easier but	on the other hand,	it makes it less personal. there
2	81B.txt	that enhance communication among people separated by distance. The internet,	on the other hand,	has not only become a quick
3	118B.txt	their time on phones or photography or chat with friends,	on the other hand,	the technology deprived many of those
4	14B.txt	e-mail in peace, agriculture, industrial missions, and so on.	On the other hand,	the telephones and e-mails have
5	15B.txt	through the time it will make you feel less personal.	On the other hand,	using emails for work place is
6	27B.txt	each book to have the information they're looking about.	On the other hand,	a lot of websites gives the
7	33B.txt	appointment, arrange meetings, transfer money .... etc. and save more time	On the other hand,	Internet is a good way of

Figure 2: Examples of "on the other hand".

4.3.1.2 Learner Errors and Learner English Corpus

When comparing learners' performance across proficiency groups or native speakers, error analysis can help researchers with signalling developmental stages. Meanwhile, learner corpus can provide traceable contexts when researchers want to analyse errors based upon certain hypotheses. QCAW was not annotated for learner errors. This makes it open to hypothesis-specific studies in the future, and researchers can develop their error annotation scheme.

The notion of learner errors can be traced back to as early as Corder's work in 1967. From then on, many studies have discussed learner errors from various perspectives, for example, parts of speech, syntax, morphology, register, appropriateness, cross-corpora … etc. Following a thorough examination of previous studies in learner errors and learner corpus research (e.g., Corder, 1967; Cowan, Choi, and Kim, 2003; Dobrić 2023; Dobrić and Sigott 2014; Ellis and Barkhuizen, 2005; Granger 1996; Lennon 1991; McEnery et al. 2019; Paquot 2008; Satake 2020; Valero Garcés 1997; Yoo and Shin 2019; Yoon and Jo 2014), the working definition of learner errors in this paper is: Errors are identified when learners use a word or an expression to denote a different meaning from native speakers, which could cause misinterpretation, ambiguity, or illogical statement. Learner errors are considered to be a necessary part of acquisition developed by learners in the process of acquiring a second language. They are erroneous outputs which learners' native-speaker counterparts would not produce.

In addition, from the perspective of linguistic forms, errors can be divided into errors of form, grammar, lexico-grammar, lexis, word redundant, word missing, word order, punctuation, style, and infelicities. While based on the causes of errors, learner errors may be induced by L1 specific factors or other factors such as intralingual factors, developmental factors, teaching-induced factors or communication strategies (Cowan, Choi, and Kim, 2003; Cross and Papp, 2008; James 1980). When errors are induced by learners' L1, they are defined as interlingual errors. Meanwhile, when induced by other factors regardless of learners' L1s, they are defined as intralingual errors, including over-generalisation errors, ignorance of rule restrictions, incomplete application of rules, false concepts hypothesised, and other unique types identified by researchers.

The example below highlights how to use QCAW to illustrate what it can offer when interpreting learner corpus data from the angle of learner errors.

(7) People that agree with the usage of technology in studying say that technology has no limits which means that it is really rare for someone to not be able to find results of what they are searching for. As well as no limitation of books and studying resources, hundreds of results appear when someone searches for a specific subject. Secondly, technology is considered an open education. [142B]

"as well as" is used to mention another item or point connected with the subject discussed in a sentence as an addition to the subject. It usually emphasises the expression preceding "as well as" than the one following it. In example (7), on the surface, "as well as" was used by the writer to add "no limitation of books and studying resources" to the context as supplementary information. However, according to the context, "as well as" did not make the argument more convincing or provide more detailed information or explanation by linking "no limitation of books and studying resources" with the sentence. This clause did not function as added supplementary information to the sentence. The meaning of "as well as" is used incorrectly in this case.

However, most students correctly used "as well as" as an addition marker, as shown in Figure 3. This indicated that cases like example (7) are more likely to be individual-specific.

(8) Recent studies and professors experiences' prove that students who use mobiles in classroom are not attentive or in the same level as disciplined ones. Therefore, a professor decides to ban using phones in class indirectly; thus he banned all electronic devices usage at all, such as laptops, except with disability students. The main problem behind banning the smartphone is the attention; students who use their smartphones lose half or more of their attention and focusing in class. They do not listen carefully nor taking notes. [204B]

	File	Left Context	Hit	Right Context
1	25B.txt	this case, the accessibility of information and remoteness in studying,	as well as	the availability of audiobooks and online educational
2	25B.txt	the research of Robert B. Kvavik, access to the Internet,	as well as	the availability of technological equipment like laptops
3	41B.txt	reduce the value of the teacher, but it helps students	as well as	the teacher in the educational process. It
4	53B.txt	intimate relationships. Means of communication do not know the spatial	as well as	the temporal obstacles; they can communicate and
5	117B.txt	development of the intellectual and intellectual abilities of the student,	as well as	the refinement of students. I agree with
6	124B.txt	networking sites is the ease of editing on their pages,	as well as	the freedom to add content that expresses
7	143B.txt	necessary as the world is developing at a fast rate,	as well as	the skills, and students gain these skills
8	162B.txt	using new technologies, such as tablets, digital cameras and computers.	as well as	the learner as he became using advanced
9	162B.txt	the shyest students a space to share, learn, express themselves	as well as	the rest of the students, and save
10	171B.txt	s more, students must be proficient in writing and reading,	as well as	the importance of teamwork. But, technology will

Figure 3: Example of "as well as".

"therefore" is used to introduce a logical result or conclusion of a fact or reason mentioned previously in the context. Similarly, the information following "thus" is expressed as a result of something mentioned in the immediate context. The writer of (8) wanted to establish a cause-and-effect relationship by using "therefore" and "thus" in this example, a declarative sentence which described a fact or phenomenon to serve as an example of the argument. The clauses were arranged sequentially. "Therefore" and "Thus" should not be used.

Examples (7) and (8) were noticed when we investigated transition markers (Ahmed et al. 2023) in the process of annotation. They were not as easily identified as word missing or inflectional errors. It is noteworthy that, first, high-proficiency participants found the above two examples in writing. Learners at this proficiency level do not usually make errors, such as incorrect word forms. However, infelicities do occur in their writing, which could imply that errors like (7) and (8) are more likely to occur when learners' writing performance achieves an advanced level. Thus, they can signal the stages of learners' writing development. The second point is that these two examples may enlighten researchers to use a hypothesis-driven approach to investigate learner errors in argumentative writing.

Apart from errors like examples (7) and (8), we found some typical and prevalent errors in learning English easily, as shown in the underlined words in examples (9) and (10).

(9) It is highly debates issue to determine whether or not email and telephone has made communication between people less personal. [220B]

(10) Furthermore, technology has make the communication between the schools management and parents easier than before and more quickly than parents attending, successful positive communication is essential to the success of the educational process. [127B]

Errors may occur in writing by proficient and less proficient learners. Our examples provide additional support for analysing writings by L1 specific bilingual speakers. For example, the activation of the L1 influenced word forms and word sequences or the impact of errors on cross-language interaction.

4.3.1.3 Learner Corpus Data and Writing Assessment

Besides investigating variability and errors, learner corpus has been proven highly significant in writing assessment. QCAW collected texts from L1 Arabic and L2 English argumentative writing tasks, which reflected the nature of task-based first/second language learning and teaching. This design focused on language performance and acquisition and promoting in-class language learning. QCAW is expected to contribute to developing assessments in task-based L1 Arabic and L2 English writing contexts from two aspects.

First, language use represented by learner corpus is not the same as in the real world. The reason is that texts collected by a learner corpus, no matter whether a corpus of spoken or written language, is usually attained in an environment that induces learners to produce certain types of language use, and they are meaning-focused (Müller-Hartmann and Schocker-von Ditfurth 2011; Skehan 1998). Therefore, a corpus containing written texts by learners from a specific L1 background will prove that some unique features can be used in writing assessment. For instance, many instances of long sentences were found in QCAW. They are much longer than the average sentence length in the writing of native speakers, which could be a result of L1 Arabic influence. For example, we even found a sentence that contained 97 words. This feature could help writing assessors determine the complexity and coherence of L2 English writing by L1 Arabic speakers.

Second, a more detailed assessment of variables which drew less attention in this area should be considered in terms of methodology. For example, QCAW collected texts from male and female participants. Gender, as a potential variable, can be combined with other variables, such as lexical complexity or proficiency level, to benefit future research on writing assessment.

5 Conclusions

The Qatari Corpus of Argumentative Writing (QCAW) is the first in the Middle East and North Africa (MENA) region. The same Qatari students wrote two argumentative essays in L1 Arabic and L2 English on different topics. The current research presented how QCAW was built in terms of its methods, procedures, and challenges. In this concluding section, we will highlight some important implications of QCAW as a learner corpora and provide some suggestions for further research. Having detailed the methodological rigour involved in constructing the QCAW corpus through its design, development and annotation stages, the implications emanating from this resource will now be discussed.

5.1 Implications of QCAW as a Learner Corpora

Building learner corpora has important implications for language teaching, learning, and linguistic research. One of the main implications of building learner corpora is that it allows for a more detailed and accurate understanding of the language learning process. For example, analysing the errors and patterns of usage in learner corpora can enable researchers to identify common problems that learners face (Blagoeva 2004). In addition, using learner corpora can help increase learners' understanding of usage patterns, reduce collocation errors, and enhance

students' autonomy and self-correction skills (Smirnova 2017). This can apply to QCAW, where researchers can identify common writing problems that Qatari university students encounter in L1 Arabic and L2 English and develop tailored teaching materials accordingly.

Another implication of building learner corpora is that teachers and researchers can develop more authentic and relevant language materials and assessments (Chen 2011). For example, by analysing learners' written language in QCAW, developers of L1 Arabic and L2 English materials can create texts and tasks more appropriate for learners' abilities and interests.

Another important implication of building learner corpora is the ability to evaluate language teaching materials and methods (Man and Chau 2019). Researchers can evaluate the teaching materials and methods by comparing the writing produced by Qatari L1 Arabic and L2 English learners with the language presented in teaching materials.

Moreover, considering learners' corpora's strengths and weaknesses can ensure they are successfully integrated into the language classes (Kaltenböck and Mehlmauer-Larcher 2005). One form of this successful integration could be teachers' provision of focused feedback on their students' mistakes (Crosthwaite 2017). L1 Arabic writing instructors and L2 English writing instructors in the Arab world, in general, and Qatar, in particular, can incorporate QCAW in their teaching and feedback to enrich students' learning experiences.

In addition to the implications mentioned above, building learner corpora allows for examining the lexical development of second language learners (Berger, Crossley, and Kyle 2019). It helps develop their collocational competence (Li 2017). This applies to our QCAW, where teachers and researchers can analyse the students' lexical complexity and develop students' collocational competence in their L1 Arabic and L2 English writing and can, therefore, provide insightful implications to the teaching and learning of vocabulary acquisition.

Furthermore, building learner corpora also allows for examining second-language learners' discourse markers and organisation. Studies have used learner corpora to investigate discourse markers (Gilquin 2016) and discourse organisation (e.g., Swales 1990) in second language learners. Consequently, QCAW can be a great opportunity for L1 Arabic and L2 English writing instructors and researchers to analyse the patterns used in discourse markers and discourse organisation, giving insightful implications to the teaching and assessment of L1 Arabic and L2 English writing.

Moreover, building learner corpora can provide insights into the pragmatic development of second language learners in different genres or contexts in characterising their use (Vaughan and Clancy 2013). Pragmatic features such as hedging (e.g., Hyland 1998) in second language learners can be analysed in our QCAW in L2

English, providing an insight into whether they are appropriately or inappropriately used.

Furthermore, building learner corpora can provide insights into the writing development of second language learners. Studies have used learner corpora to investigate cohesive devices (e.g., Biber and Finegan 1989; Hyland 1998) and text organisation (e.g., Bhatia 1993; Swales 1990) in second language learners. Using QCAW can help students identify how different transition markers and paragraph development in L1 Arabic and L2 English were used by analysing patterns in the corpus.

Besides, building learner corpora can provide insights into the interlanguage development of second language learners. For example, in Hernández and Paredes' study (2005), students, in their corpus-based research, overused highly technical and general vocabulary in their writing, which confirms the interlanguage factor. Similarly, researchers can analyse the QCAW and find L2 English writers' patterns that could be attributed to L1 Arabic interlanguage or vice versa.

In conclusion, building learner corpora has numerous implications for language teaching, learning, and linguistic research. Learner corpora provide insights into the language learning process, allow for the development of more authentic and relevant language materials and assessments, enable the evaluation of language teaching materials and methods, and help develop learners' linguistic competence. Furthermore, building learner corpora can provide insights into second language learners' writing, pragmatic, and interlanguage development. Thus, creating and analysing learner corpora, such as the Qatar Corpus of Argumentative Writing (QCAW), can significantly improve language education and research in L1 Arabic and L2 English.

5.2 Suggestions for Further Research

The Qatari Corpus of Argumentative Writing (QCAW) can be a valuable resource for the following areas of future research.

1. **Contrastive Rhetoric:** The corpus can investigate the rhetorical and linguistic differences between Arabic and English argumentative writing, particularly in terms of organisation, coherence, and argumentation strategies. This can provide insights into the cultural and linguistic factors that shape argumentation in different contexts and inform the development of cross-cultural communication skills.
2. **L1 Arabic/L2 English Writing Research**: The corpus can be used to investigate the development of argumentative writing skills among Arabic-speaking students in L1 Arabic and L2 English. This can involve exploring the influence of L1 transfer, the role of feedback and instruction, or the impact of individual differences on writing proficiency.

3. **Studying Language Transfer:** The corpus can be used to study the influence of students' first language (L1 Arabic) on their second language (L2 English) argumentative writing and vice versa. This can help identify common errors and challenges that Arabic-speaking students face when writing in English and inform the design of more effective writing instruction and feedback.
4. **Automated Essay Scoring:** The corpus can be used to develop and evaluate automated scoring systems for argumentative essays written by Arabic-speaking students in L1 Arabic and L2 English. This can involve exploring the performance of different machine learning algorithms, features, or scoring models or investigating the impact of factors such as prompt complexity or genre.
5. **Writing Pedagogy:** The corpus can inform the design and evaluation of writing pedagogy for Arabic-speaking students learning argumentative writing in L1 Arabic and L2 English. This can involve exploring the effectiveness of different instructional approaches, feedback strategies, and writing tasks or investigating the impact of individual differences on writing proficiency.
6. **Cross-linguistic Research:** The corpus could be used to conduct cross-linguistic research to analyse linguistic aspects (.e.g. metadiscourse and voice; syntactic complexity, use of collocations, grammatical complexity, style of writing, cohesion, coherence, mechanics of writing) in L1 Arabic and L2 English. The corpus can explore linguistic phenomena in L1 Arabic and L2 English argumentative writing, including discourse markers, rhetorical devices, or lexical and grammatical features. This can involve investigating the role of these phenomena in argumentative writing, their relationship to other linguistic and rhetorical factors, or their impact on text quality or persuasiveness.

Acknowledgement: This paper is the outcome of a funded research grant by the Qatar National Research Fund (QNRF), Doha, Qatar. (NPRP Grant No. NPRP11S-1112-170006). Qatar National Library (QNL), Doha, Qatar, provided open-access funding for this article.

Appendix 1

Student and Instructor Responses to Topics

No.	Essay Question	Instructors	Students
1.	Some students prefer to study alone. Others prefer to study with a group of students. Which do you prefer? Use specific reasons and examples to support your answer.	66.7%	50%
2.	Do you agree or disagree with the following statement? Watching television is bad for children. Use specific details and examples to support your answer.	33.3%	64.70%
3.	Do you agree or disagree with the following statement? Children should begin learning a foreign language (e.g. English) as soon as they start school. Use specific reasons and examples to support your position.	83.3%	61.70%
4.	Do you agree or disagree with the following statement? Telephones and emails have made communication between people less personal. Use specific reasons and examples to support your opinion.	66.6%	64.70%
5.	Many teachers assign homework to students every day. Do you think that daily homework is necessary for students? Use specific reasons and details to support your answer.	50%	58.80%
6.	Do you agree or disagree with the following statement? Face-to-face communication is better than other types of communication, such as letters, email, or telephone calls. Use specific reasons and details to support your answer.	50%	52.90%
7.	Do you agree or disagree with the following statement? With the help of technology, students nowadays can learn more information and learn it more quickly. Use specific reasons and examples to support your answer.	83.40%	73.50%
8.	Is it better to enjoy your money when you earn it or is it better to save your money for some time in the future? Use specific reasons and examples to support your opinion.	83.40%	41.20%
9.	Some people think that they can learn better by themselves than with a teacher. Others think that it is always better to have a teacher. Which do you prefer? Use specific reasons to develop your essay.	83.30%	44.10%
10.	It has been said, "Not everything that is learned is contained in books." Compare and contrast knowledge gained from experience with knowledge gained from books. In your opinion, which source is more important? Why?	66.6	44%

Appendix 2

Task and Writing Prompt Instructions

Dear Student,

Please write an argumentative essay based on your personal knowledge and experience about the given topic below in *no less than 500 words* about the writing topic assigned to you. In your writing, make sure that you:

- have a clear thesis statement supported by relevant evidence,
- establish a clear relevance of the arguments to the essay topic,
- develop critical thoughts by presenting opposing views, support them with evidence, and make your position clear on the issue
- spend NO MORE THAN 50 minutes in full doing the task.

Prompt 1:
Do you agree or disagree with the following statement? Telephones and emails have made communication between people less personal. Use specific reasons and examples to support your opinion.

Prompt 2:
Do you agree or disagree with the following statement? With the help of technology, students nowadays can learn more information and learn it more quickly. Use specific reasons and examples to support your answer.

Appendix 3

Argumentative Writing Rubric in English

Deanship of General Studies - Core Curriculum Program – Qatar University

Argumentative Writing Rubric

Criterion	Excellent	Good	Acceptable	Poor
Introduction (Proposition) **(1 mark)**	A student writes an attractive introduction to his/her essay in an organised way, referring to an excellent thesis statement at the end of the introduction. (1 mark)	A student writes a good introduction to his/her essay in a comprehensive way, referring to a good thesis statement at the end of the introduction. (0.5 mark)	A student writes an unattractive introduction to his/her essay in an incomprehensive way, referring to an incomplete thesis statement at the end of the introduction. (0.25 mark)	A student does not know how to introduce his/her essay in an organised way, lacking a thesis statement at the end of the introduction. (0.0 mark)
Presentation of the main ideas (Explanation of Discussed Issues) **(2 marks)**	A student presents the main ideas very clearly and comprehensively, providing all the specific details and examples needed. (2 marks)	A student presents the main ideas clearly and adequately, providing some specific details and examples needed. (1.5 marks)	A student presents the main ideas clearly to some extent, providing few specific details and examples needed. (1 mark)	A student presents the main ideas briefly and superficially, lacking the specific details and examples needed. (0.5 mark)
Student's Critical Response (Opinion) **(4 marks)**	A student writes his/her stance to the essay question prompt (i.e. agreeing/disagreeing) providing a justified and comprehensive evidence. (4 marks)	A student writes his/her stance to the essay question prompt (i.e. agreeing/ disagreeing) providing a justified and an acceptable evidence. (2 marks)	A student writes his/her stance to the essay question prompt (i.e. agreeing or disagreeing) lacking a justified evidence. (1 mark)	A student writes his/her stance to the essay question prompt (i.e. agreeing or disagreeing) and does not provide any evidence. (0.5 mark)
Resources and Supporting Evidence (2 marks)	A student supports his/her critical stance by citing three or more resources or examples in text. (2 marks)	A student supports his/her critical stance by citing two or more resources or examples in text. (1.5 marks)	A student supports his/her critical stance by citing at least one resource or example in text. (1 mark)	A student uses an unknown resource or example or lacks resources or examples to support his/her critical stance in text. (0.5 mark)
Conclusion (Deductions and Relevant Results) **(1 mark)**	A student concludes his/her essay well by summarizing all ideas discussed in the essay and the lessons learnt. (1 mark)	A student concludes his/her essay by briefly summarizing the ideas discussed in the essay and the lessons learnt. (0.50 mark)	A student concludes his/her essay well by partially summarizing the ideas discussed in the essay, and the lessons learnt. (0.25 mark)	A student concludes his/her essay badly by neither summarizing the ideas discussed in the essay nor referring to the lessons learnt. (0.0 mark)

References

Abu-Rabia, S., and J. Awwad. 2004. "Morphological Structures in Visual Word Recognition: The Case of Arabic." *Journal of Research in Reading* 27 (3): 321—36.

Ädel, A. 2015. "Variability in Learner Corpora." In *Cambridge Handbook of Learner Corpus Research*, edited by S. Granger, G. Gilquin, and F. Meunier, 379—400. UK: Cambridge University Press.

Ahmed, A. 2010a. "Contextual Challenges to Egyptian Students' Writing Development." *International Journal of Arts and Sciences* 3 (14): 503—22.

Ahmed, A. 2010b. "The EFL Essay Writing Difficulties of Egyptian Student Teachers of English: Implications for Essay Writing Curriculum and Instruction." Unpublished PhD Thesis, Graduate School of Education, University of Exeter.

Ahmed, A. M., X. Zhang, L. M. Rezk, and W. S. Pearson. 2023. “Transition Markers in Qatari University Students’ Argumentative Writing: Across-Linguistic Analysis of L1 Arabic and L2 English.” *Ampersand* 10: 100110.

Ahmed, A., and D. Myhill. 2016. “The Impact of the Socio-Cultural Context on L2 English Writing of Egyptian University Students.” *Learning, Culture and Social Interaction* 11: 117—29.

Ahmed, A., D. Myhill, E. Abdollahzadeh, L. McCallum, W. Zaghouani, L. Rezk, A. Jrad, and X. Zhang. 2022. *The Qatari Corpus of Argumentative Writing (QCAW). LDC2022T04. Web Download*. Philadelphia: Linguistic Data Consortium, USA. https://catalog.ldc.upenn.edu/LDC2022T04.

Al-Jarf, R. 2013. “Enhancing Freshman Students’ Performance with Online Reading and Writing Activities.” In *Conference Proceedings of eLearning and Software for Education «(eLSE)*, Vol. 9, 2nd ed., 524—30. Carol I National Defence University Publishing House.

Al-Jarf, R. 2022. “English Transliteration of Arabic Personal Names with the Definite Article {al-} on Facebook.” *British Journal of Applied Linguistics* 2 (2): 24—31.

Al-Khatib, M. A. 2000. “The Arab World: Language and Cultural Issues.” *Language Culture and Curriculum* 13 (2): 121—5.

Al-Khatib, M.A. 1988. “Sociolinguistic Change in an Expanding Urban Context: A Case Study of Irbid City, Jordan.” Unpublished PhD thesis, University of Durham.

Al-Khatib, M.A. 1995. “The Impact of Sex on Linguistic Accommodation: A Case Study of Amman Radio Phone-In Program.” *Multilingua* 14 (2): 133—50.

Al-Momani, I. M. 2011. “The Syntax of Sentential Negation in Jordanian Arabic.” *Theory and Practice in Language Studies* 1 (5): 482—96.

Alqahtni, H. M. 2014. “The Structure and Context of Idiomatic Expressions in the Saudi Press.” Doctoral diss., University of Leeds.

Berger, C. M., S. A. Crossley, and K. Kyle. 2019. “Using Native-Speaker Psycholinguistic Norms to Predict Lexical Proficiency and Development in Second-Language Production.” *Applied Linguistics* 40 (1): 22—42.

Bhatia, V. K. 1993. *Analysing Genre: Language Use in Professional Settings*. London and New York: Longman.

Biber, D., and E. Finegan. 1989. “Styles of Stance in English: Lexical and Grammatical Marking of Evidentiality and Affect.” *Text-Interdisciplinary Journal for the Study of Discourse* 9 (1): 93—124.

Blagoeva, R. 2004. “Demonstrative Reference as a Cohesive Device in Advanced Learner Writing: A Corpus-Based Study.” In *Advances in Corpus Linguistics*, 297—307. Göteborg: Brill.

Caines, A., M. McCarthy, and A. O’Keeffe. 2016. “Spoken Language Corpora and Pedagogical Applications.” In *The Routledge Handbook of Language Learning and Technology*, 348—61. Abingdon: Routledge.

Callies, M. 2013. “Advancing the Research Agenda of Interlanguage Pragmatics: The Role of Learner Corpora.” In *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, 9—36. New York City: Springer International Publishing.

Chen, A., and F. C. Gey. 2002. “Building an Arabic Stemmer for Information Retrieval.” In *TREC*, 2002, 631—9. USA: National Institute of Standards & Technology.

Chen, H. J. 2011. “Developing and Evaluating a Web-Based Collocation Retrieval Tool for EFL Students and Teachers.” *Computer Assisted Language Learning* 24 (1): 59—76.

Corder, S. 1967. “The Significance of Learners’ Errors.” *International Review of Applied Linguistics* 54: 161—70.

Cowan, R., H. Choi, and D. Kim. 2003. “Four Questions for Error Diagnosis and Correction in CALL.” *CALICO Journal* 20 (3): 451—63.

Coxhead, A. 2000. “A New Academic Word List.” *Tesol Quarterly* 34 (2): 213—38.
Crompton, P. 2011. “Article Errors in the English Writing of Advanced L1 Arabic Learners: The Role of Transfer.” *Asian EFL Journal* 50 (1): 4—35.
Cross, J., and S. Papp. 2008. “Creativity in the Use of Verb+ Noun Combinations by Chinese Learners of English.” In *Linking Up Contrastive and Learner Corpus Research*, 55—81. Oxford: Oxford University Press.
Crosthwaite, P. 2017. “Retesting the Limits of Data-Driven Learning: Feedback and Error Correction.” *Computer Assisted Language Learning* 30 (6): 447—73.
Dashtestani, R., and N. Stojkovic. 2016. “The Use of Technology in English for Specific Purposes (ESP) Instruction: A Literature Review.” *Journal of Teaching English for Specific and Academic Purposes* 3 (3): 435—56.
Dobrić, N. 2023. “Identifying Errors in a Learner Corpus—The Two Stages of Error Location vs. Error Description and Consequences for Measuring and Reporting Inter-annotator Agreement.” *Applied Corpus Linguistics* 3 (1): 100039.
Dobrić, N., and G. Sigott. 2014. “Towards an Error Taxonomy for Student Writing.” *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 19 (2): 111—8.
Ellis, R., and G. Barkhuizen. 2005. *Analysing Learner Language*. Oxford: Oxford University Press.
Gamon, M., M. Chodorow, C. Leacock, J. Tetreault, N. Ballier, A. Dıaz-Negrillo, and P. Thompson. 2013. “Using Learner Corpora for Automatic Error Detection and Correction.” In *Automatic Treatment and Analysis of Learner Corpus Data*, 127—50. The Netherlands: John Benjamins.
Gilquin, G. 2016. “Discourse Markers in L2 English.” In *New Approaches to English Linguistics: Building Bridges*, 213—49. Amsterdam, The Netherlands: John Benjamins.
Gilquin, G., and S. Granger. 2015. “Learner Language.” In *The Cambridge Handbook of English Corpus Linguistics*, edited by D. Biber, and R. Reppen, 418—35. Cambridge University Press.
Gilquin, G., S. Granger, and M. Paquot. 2007. “Learner Corpora: The Missing Link in EAP Pedagogy.” *Journal of English for Academic Purposes* 6 (4): 319—35.
Granger, S. 1996. “From CA to CIA and Back: An Integrated Approach to Computerised Bilingual and Learner Corpora.” In *Languages in Contrast. Papers from a Symposium on Text-Based Cross-Linguistic Studies*, edited by K. Aijmer, B. Altenberg, and M. Johansson. Lund: Lund University Press, 37—51.
Granger, S. 2003. “Error-tagged Learner Corpora and CALL: A Promising Synergy.” *CALICO Journal* 20: 465—80.
Granger, S. 2009. “The Contribution of Learner Corpora to Second Language Acquisition and Foreign Language Teaching.” *Corpora and Language Teaching* 33: 13—32.
Granger, S. 2011. “How to Use Foreign and Second Language Learner Corpora.” In *Research Methods in Second Language Acquisition: A Practical Guide*, 5—29.
Granger, S., G. Gilquin, and M. Fanny. 2015. “Introduction: Learner Corpus Research — Past, Present and Future.” In *The Cambridge Handbook of Learner Corpus Research*, edited by S. Granger, G. Gilquin, and F. Meunier, 1—5. Cambridge University Press.
Gries, S. T., and S. Wulff. 2020. “Examining Individual Variation in Learner Production Data: A Few Programmatic Pointers for Corpus-Based Analyses Using the Example of Adverbial Clause Ordering.” *Applied PsychoLinguistics* 42 (2): 279—99.
Habash, N., and D. M. Palfreyman. 2023. “ZAEBUC Design and Annotation.” In *Bilingual Writers and Corpus Analysis*, edited by D. M. Palfreyman, and N. Habash, 19—51. Routledge.
Hamed, O., and T. Zesch. 2017. “A Survey and Comparative Study of Arabic Diacritization Tools.” *Journal for Language Technology and Computational Linguistics* 32 (1): 27—47.

Hendriks, H., ed. 2005. *The Structure of Learner Varieties*. Berlin: Mouton-de Gruyter.
Hernández, P. S., and P. F. P. Paredes. 2005. "Examining English for Academic Purposes Students' Vocabulary Output: Corpus-Aided Analysis and Learner Corpora." *Revista Espanola de Linguistica Aplicada* Extra 1 (1): 201—12.
Hertel, T. J. 2003. "Lexical and Discourse Factors in the Second Language Acquisition of Spanish Word Order." *Second Language Research* 19 (4): 273—304.
Hyland, K. 1998. "Hedging in Academic Writing and EAP Textbooks." *English for Specific Purposes* 17 (1): 3—25.
Hyland, K. 2010. "Metadiscourse: Mapping Interactions in Academic Writing." *Nordic Journal of English Studies* 9 (2): 125—43.
James, C. 1980. *Contrastive Analysis*. London: Longman.
Jarvis, S. 2000. "Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon." *Language Learning* 50 (2): 245—309.
Kaltenböck, G., and B. Mehlmauer-Larcher. 2005. "Computer Corpora and the Language Classroom: On the Potential and Limitations of Computer Corpora in Language Teaching." *ReCALL* 17 (1): 65—84.
Kayaoglu, M. 2013. "The Use of Corpus for Close Synonyms." *Journal of Language and Linguistic Studies* 9 (1): 128—44.
Kaye, A. S. 2017. "Arabic." In *The World's Major Languages*, 576—93. London: Routledge.
Khorsheed, M. S. 2002. "Off-line Arabic Character Recognition—A Review." *Pattern Analysis & Applications* 5: 31—45.
Khuwaileh, A. A., and A. A. Shoumali. 2000. "Writing Errors: A Study of the Writing Ability of Arab Learners of Academic English and Arabic at University." *Language Culture and Curriculum* 13 (2): 174—83.
Kirmizi, O., and Karci, B. 2017. An Investigation of Turkish Higher Education EFL Learners' Linguistic and Lexical Errors. *Educational Process: International Journal* 6(4): 35.
Kusters, W. 2003. *Linguistic Complexity*. The Netherlands: Netherlands Graduate School of Linguistics.
Lennon, P. 1991. "Error: Some Problems of Definition, Identification, and Distinction." *Applied Linguistics* 12 (2): 180—96.
Li, S. 2017. "Using Corpora to Develop Learners' Collocational Competence." *Language, Learning and Technology* 21 (3): 153—71.
Man, D., and M. H. Chau. 2019. "Learning to Evaluate through That-Clauses: Evidence from a Longitudinal Learner Corpus." *Journal of English for Academic Purposes* 37: 22—33.
McEnery, T., V. Brezina, D. Gablasova, and J. Banerjee. 2019. "Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyse Language Use." *Annual Review of Applied Linguistics* 39: 74—92.
Mollin, S. 2006. *Euro-English: Assessing Variety Status*. Tübingen: Gunter Narr Verlag.
Mourssi, A. 2013. "Cross-linguisticInfluence of L1 (Arabic) in Acquiring Linguistic Items of L2 (English): An Empirical Study in the Context of Arab Learners of English as Undergraduate Learners." *Theory and Practice in Language Studies* 3 (3): 397—403.
Müller-Hartmann, A., and M. Schocker-von Ditfurth. 2011. *Introduction to English Language Teaching: Optimise Your Exam Preparation*. Stuttgart: Klett Lerntraining.
Naz, S., A. I. Umar, S. H. Shirazi, S. B. Ahmed, M. I. Razzak, and I. Siddiqi. 2016. "Segmentation Techniques for Recognition of Arabic-like Scripts: A Comprehensive Survey." *Education and Information Technologies* 21: 1225—41.

Okamoto, K. 2010. “Incorporating Corpora into English Language Teaching for Undergraduate Computer Science and Engineering Students with Limited Proficiency.” In *2010 IEEE International Professional Communication Conference*, 152—6. IEEE. Enschede, Netherlands.

Olsen, S. 1999. “Errors and Compensatory Strategies: A Study of Grammar and Vocabulary in Texts Written by Norwegian Learners of English.” *System* 27 (2): 191—205.

Paquot, M. and Fairon, C. 2006. “Investigating L1-Induced Learner Variability: Using the Web as a Source of L1 Comparable Data.” In *Paper Presented at the International Computer Archive of Modern and Medieval English (ICAME) Conference (Variation, Contacts and Change), University of Helsinki, 24—28 May 2006*.

Paquot, M. 2008. “Exemplification in Learner Writing: A Cross-Linguistic Perspective.” In *Phraseology in Foreign Language Learning and Teaching*, edited by F. Meunier, and S. Granger, 101—19. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Pendar, N., and C. A. Chapelle. 2008. “Investigating the Promise of Learner Corpora: Methodological Issues.” *CALICO Journal* 25 (2): 189—206.

Phoocharoensil, S. 2013. “Cross-linguistic Influence: Its Impact on L2 English Collocation Production.” *English Language Teaching* 6 (1): 1—10.

Regan, V. 2013. “Variation.” In *The Cambridge Handbook of Second Language Acquisition*, edited by J. Herschensohn, and M. Young-Scholten, 272—91. Cambridge; New York: Cambridge University Press.

Saiegh-Haddad, E., and R. Henkin-Roitfarb. 2014. “The Structure of Arabic Language and Orthography.” In *Handbook of Arabic Literacy: Insights and Perspectives*, 3—28. Dordrecht, Netherlands: Springer.

Salloum, S., T. Gaber, S. Vadera, and K. Shaalan. 2023. “A New English/Arabic Parallel Corpus for Phishing Emails.” *ACM Transactions on Asian and Low-Resource Language Information Processing* 22 (7): 1—17.

Satake, Y. 2020. “How Error Types Affect the Accuracy of L2 Error Correction with Corpus Use.” *Journal of Second Language Writing* 50: 100757.

Sawalha, M., and E. Atwell. 2013. “A Standard Tag Set Expounding Traditional Morphological Features for Arabic Language Part-Of-Speech Tagging.” *Word Structure* 6 (1): 43—99.

Shirato, J., and P. Stapleton. 2007. “Comparing English Vocabulary in a Spoken Learner Corpus with a Native Speaker Corpus: Pedagogical Implications Arising from an Empirical Study in Japan.” *Language Teaching Research* 11 (4): 393—412.

Skehan, P. 1998. *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Smirnova, E. A. 2017. “Using Corpora in EFL Classrooms: The Case Study of IELTS Preparation.” *RELC Journal* 48 (3): 302—10.

Swales, J. M. 1990. *Genre Analysis: English in Academic and Research Settings*. USA: Cambridge University Press.

Valero Garcés, C. 1997. “The Interlanguage of Spanish Students Beginning English Philology.” *GRETA* 5 (2): 74—8.

Vaughan, E., and B. Clancy. 2013. “Small Corpora and Pragmatics.” In *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, 53—73. Dordrecht, Netherlands: Springer.

Vyatkina, N. 2013. “Specific Syntactic Complexity: Developmental Profiling of Individuals Based on an Annotated Learner Corpus.” *The Modern Language Journal* 97 (S1): 11—30.

Wulff, S., and S. Gries. 2021. “Exploring Individual Variation in Learner Corpus Research: Methodological Suggestions.” In *Learner Corpus Research Meets Second Language Acquisition*, edited by B. Le Bruyn, and M. Paquot, 191—213. Cambridge: Cambridge University Press.

Yoo, I. W., and Y. K. Shin. 2019. “Determiner Use in English Quantificational Expressions: A Corpus-Based Study.” *Tesol Quarterly* 54: 90—117.

Yoon, H., and J. W. Jo. 2014. “Direct and Indirect Access to Corpora: An Exploratory Case Study Comparing Students’ Error Correction and Learning Strategy Use in L2 Writing.” *Language, Learning and Technology* 18 (1): 96—117.

Yoon, S. 2020. “The Learner Corpora of Spoken English: What Has Been Done and what Should Be Done?” *Language Research* 56 (1): 29—51.