QCAW 1.0: Building a Qatari Corpus of Student Argumentative Writing

Wajdi Zaghouani [1], Abdelhamid Ahmed [2], Xiao Zhang [3] and Lameya Rezk [2]

[1] Hamad Bin Khalifa University
[2] Qatar University
[3] Xi'an International Studies University

wzaghouani@hbku.edu.qa , aha202@qu.edu.qa , 2750908792@qq.com , lmh207@yahoo.com

Abstract

This paper 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. The Metadata 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 paper 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.

Keywords: Arabic, Argumentative Writing, Corpus Building, Bilingual Writer Corpus

1. Introduction

Learner corpora are authentic language data produced by individuals learning their first or second language (Granger et al., 2015; Gilquin et al., 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 multi-word clusters that learner's 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 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 et al. (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 (Ahmed et al., 2023).

The main contributions of this work are:

- The development of the Qatari Corpus of Argumentative Writing (QCAW), the first publicly available parallel corpus of L1 Arabic and L2 English essays.
- The corpus provides a valuable new resource for research in contrastive rhetoric, automated writing evaluation, error analysis, and pedagogy for Arabic learners.
- 390 argumentative essays written by 195 Qatari students in both Arabic and English on two different topics and rated for overall writing quality and voice allow for multifaceted analyses.
- Detailed metadata supports investigating effects of variables like gender and discipline.

2. Related Works

This section provides an overview of learner corpora, which are pivotal for research in second language acquisition and language processing. The various types of learner corpora, such as written and spoken corpora, enable detailed study into language learners' proficiency and development. These corpora are often characterized by parameters such as the time and scope of collection, the targeted language of learners (L2), the learners' native language (L1), the medium, and the text type, with argumentative writing and informal spoken interviews being common

LREC-COLING 2024, pages 13382–13394
20-25 May, 2024.

focuses (Granger, 2011). Additionally, Arabic-English bilingual corpora are crucial for examining the nuances of L1 Arabic and L2 English writing, contributing to a deeper understanding of language learning processes and interlanguage development (Habash & Palfreyman, 2022).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 (Gilquin & Granger, 2015; Coxhead, 2000). 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 (Yoon, 2020; Caines et al.,2016). 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 et al., 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 represented text types in learner corpora are argumentative texts for writing and informal interviews for speaking.

2.2 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 & Palfreyman (2022). 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 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.

2.3 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 (Habash et al., 2007). 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 & 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 & 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 & Atwell, 2013).

Arabic written language has an inflexion system, known as declensions, which indicate the grammatical function of nouns and adjectives (Saiegh-Haddad & 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 the word (Abu-Rabia & 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 & Gey, 2002). Another characteristic of the inflexion system is that Arabic verbs have a complex conjugation system based on the person, gender, and the number of the subject (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.

2.4 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 & 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 & 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, 2010). 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, lexical, grammatical and mechanics (Ahmed, 2010), making it difficult for readers to understand the intended meaning. These problems are attributed to socio-cultural issues (Ahmed & 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, 2010).

2.5 Corpora and Annotation Frameworks

The development of the Qatari Corpus of Argumentative Writing (QCAW) is supported by a foundation of works that have advanced the annotation and analysis of Arabic corpora. Initiatives such as the Arabic Propbank (Palmer et al., 2008; Diab et al., 2008) have been instrumental in semantic role labeling, which is crucial for dissecting argumentative structures within bilingual corpora.

The Propbank has seen revisions and expansions, including annotations for Quranic Arabic (Zaghouani, Hawwari, & Diab, 2012), highlighting the evolving nature of Arabic linguistic resources. These frameworks have directly influenced the methodologies for corpus creation and annotation applied to the QCAW. Educational tools like ARET further support Arabic language learning and processing, offering insights into reading enhancement (Maamouri et al., 2012).

Crowdsourcing has emerged as a viable method for language resource annotation, including for Arabic, presenting an innovative way to gather linguistic data (Zaghouani & Dukes, 2014). This aspect of collaborative annotation is reflected in the QCAW's metadata, which details student contributors and text characteristics. Large-scale annotation frameworks for Arabic have set guidelines that have been taken into account for the QCAW, ensuring the corpus' error annotation is both rigorous and systematic (Zaghouani et al., 2014; Zaghouani et al., 2015a).

The critical survey by Zaghouani (2014) provides a comprehensive overview of freely available Arabic corpora, situating the QCAW within the wider context of available resources. The QALB shared tasks on automatic text correction for Arabic also contribute to the corpus's design, especially in enhancing the text's quality through automatic error correction techniques (Mohit et al., 2014; Rozovskaya et al., 2015).

Additionally, the QCAW's creation aligns with the efforts to build corpora specific to Qatari Arabic expressions, enriching the representation of regional language varieties (Al-Mulla & Zaghouani, 2020). A scoping review by Ahmed et al. (2022) on Arabic corpora emphasizes the importance of such specialized resources in the broader landscape of computational linguistics.

Together, these sources provide a multifaceted framework for understanding and developing bilingual corpora. The QCAW's structure and annotations are the result of cumulative advancements in Arabic linguistic resources, demonstrating the corpus's potential for facilitating research into bilingual language use and second language acquisition.

3. Methodology

This section highlights how the Qatari Corpus of Argumentative Writing (QCAW) was built, taking into consideration the essay writers, the selection of writing topics, and task completion.

3.1 Corpus Description

The Qatari Corpus of Argumentative Writing (QCAW) is a comprehensive collection consisting of 390 essays, meticulously curated to facilitate the study of argumentative discourse (Ahmed et al., 2023). Within this corpus, an equal distribution is observed between the two primary languages of instruction in the university: English and Arabic, with 195 essays composed in each language. These essays emanate from a diverse group of students enrolled in various major programs, fostering a multifaceted representation of academic perspectives. To provide a holistic perspective, each student contributed a pair of essays, one in English and one in Arabic, each addressing distinct topics. The English corpus encompasses 98,379 words, while the Arabic counterpart comprises 97,248 words, rendering a balanced linguistic distribution within the QCAW.

The QCAW demonstrates a blend of gender representation, reflecting the broader demographics of the university population. Among the 390 essays, the majority were authored by female students (n=159), with male students contributing 36 essays. This gender distribution underscores the existing gender imbalances in the academic domain, adding an additional layer of complexity to the corpus. The essays within the corpus span various grade levels, as determined by analytical writing quality ratings, with approximately 70% of English texts and 68% of Arabic texts falling into the mid-level category, achieving scores ranging from 5 to 7 out of 10. The voice ratings further reveal an insightful perspective, with the majority of texts (64-70%) clustered within the mid-level range, receiving voice scores of 2 to 3 out of 5.

For a detailed overview of the corpus composition and rating distributions, please refer to Table 2 and 3, which succinctly summarize these essential characteristics. The QCAW thus stands as a rich and diverse resource, poised to facilitate multifaceted investigations into argumentative discourse across linguistic and gender dimensions.

3.2 Topic Selection

To select appropriate writing prompts for the corpus, an initial pool of 10 argumentative essay topics were drawn from the TOEFL independent writing task prompts. A survey was conducted with 34 students and 6 instructors at the university to gather feedback on which topics would be most relevant and engaging. The top 2 topics based on this survey were chosen as the final prompts for the corpus (see Appendix 1). These prompts asked students to write an evidence-based argumentative essay in response to a question about technology's impact on education and communication.

3.3 Data Sources and Collection

Students ranged from 18 to 22 years old. They were in different years of study in different colleges. They were bilingual students whose L1 was Arabic, and L2 was English. The corpus contained more texts written by females (n= 159) than males (n= 36) because the university's female-to-male student ratio is 3 to 1. Education at the university is segregated, and instructors are assigned to teach female or male participants. The average essay length in words (Arabic 498.71, English 504.51).

Students were taking a compulsory First-Year Seminar course for undergraduate university students from diverse disciplines in the university. The course focuses on developing students' critical reading, writing, research and academic success skills. We then asked these participating students to write two argumentative essays: One in L1 Arabic and another in L2 English.

The essays comprised two topics chosen for the study through student-instructor consultation. We began by selecting ten argumentative topics from TOEFL essay writing prompts. We designed two questionnaires: one for students and the other for instructors to elicit their views on the topics of interest. Six instructors and thirty-four students responded to the questionnaire showing their preferred topics. Appendix 1 summarises the results for both instructors and students. The top two topics with the highest percentage rank from students' and instructors' perspectives were selected as the final prompts. Based on these results, the task and prompt writing instructions were designed.

Students were divided into groups A and B to ensure the corpus was balanced and representative and to mitigate task and topic effects (see Table 1). Group A completed Topic 1 in Arabic and Topic 2 in English, and Group B completed the topics in reverse order. The creation of two groups mitigated any effect caused by the writing topic.

Writing Group	N	Topic	N	Topic	Total
English Writing Group	154	1	41	2	195
Arabic Writing Group	41	2	154	1	195

Table 1. Task cross-over design

Both groups were asked to write an argumentative essay based on their knowledge and experience about the assigned topics. We piloted the selected writing topics with six students at the beginning of

September 2019. These six students were asked to report any problems with the writing task time, word count and/or the clarity of the instructions or writing topics. The pilot indicated that the conditions and writing topics were appropriate for the participating students.

3.4 Essays Annotation for Writing Quality and Voice

The English and Arabic texts went through two separate annotation or rating processes. The essays were first rated for writing quality using an analytical rubric by instructors at the concerned university. Then, they were rated for dimensions of voice using the voice rubric from Zhao (2012). The procedures for these two activities are described as follows.

English and Arabic texts were graded according to a five-category analytical rubric with a maximum possible score of 10 marks (see Appendix 3). More weight was placed on students' stance than the structural parts of the essay (e.g., introduction, developmental paragraphs, and conclusion). The analytical rubric assessed the writing quality construct under the sub-constructs: introduction, presentation of main ideas, student's stance, supporting ideas with examples or resources, and conclusion.

An initial benchmarking procedure to check grade reliability was carried out. Four annotators graded five essays, and areas of discrepancy were discussed and rectified. The same annotators rated the complete English and Arabic essays in the corpus. During this time, alignment between annotators was checked and rectified through discussions, where necessary, so each essay was not graded as more than one band score of a difference.

As for voice, we measured holistic voice salience and analytical voice. Following Zhao (2012, 2017) and Yoon (2017), a holistic voice rubric was used, which was scored out of five. These five levels are presence and centrality of ideas in the content, manner of presentation, and writer and reader presence. Four annotators were involved in the rating process to ensure the Reliability and validity of the assessments. These annotators were native speakers of Arabic; two annotators were Egyptian, and two were Tunisian. An Egyptian rater teamed up with a Tunisian one. The first two annotators assessed 98 texts, and the second two assessed 97 texts. During the rating process, annotators went through the following procedures:

3.4.1 Annotation Procedures

The annotation process commenced with annotators or annotators acquainting themselves with the rubric's fundamental structure and content. Subsequently, a norming session was conducted, during which the annotators diligently applied the voice rubric to evaluate writing samples, serving as benchmarks for the procedure. A collaborative exchange of their rating results ensued, fostering valuable discussions on the rationale behind their assessments and the interpretation of specific rubric descriptors. These deliberations played a pivotal role in achieving consensus among the annotators regarding the rubric's interpretation and its practical application to the actual writing samples. With the increasing consistency in voice ratings among the annotators and the resolution of most rating discrepancies to within a 1-point range, the rater training session culminated. Subsequently, the annotators proceeded to independently rate the writing samples for voice.

Furthermore, all Arabic and English essays underwent an assessment of their overall writing quality, scored on a 10-point scale using an analytical rubric (see Appendix 3). This rubric considered elements such as introduction, main idea presentation, utilization of textual evidence, articulation of stance, and conclusion. To ensure the standardization of the rating process, benchmark essays were employed to train four annotators and establish uniform rating standards. Throughout the comprehensive rating procedure, periodic alignment checks were diligently performed among the annotators, with any discrepancies promptly resolved through dialogue.

Additionally, each essay underwent a voice rating on a 5-point scale using Zhao's (2012) holistic voice rubric, with two native Arabic-speaking annotators assigned to each text. To maintain rating reliability, these annotators underwent training sessions to ensure a consistent understanding of the voice rubric. To further enhance the reliability of voice ratings, periodic breaks were incorporated, enabling annotators to revisit the rubric and calculate the average of the two voice scores assigned to each text.

3.4.2 General Instructions to Ensure the Annotation Reliability

Annotators were not supposed to evaluate writing quality; instead, they focused solely on assessing voice salience as defined by the rubric. They took a break at least every two hours to avoid the potential influence of fatigue on the ratings, They re-studied the rubric every time they started a new rating session to help ensure consistency across different rating sessions. Each writing sample was double-rated, and the average of the two ratings was used in the subsequent statistical analysis.

3.5 Corpus Cleaning and Preparation

In the final corpus, many texts were excluded. Four exclusion criteria were followed. Firstly, we excluded essays of less than 250 words to avoid inflated values from very short argumentative essays (McNamara, Graesser, McCarthy, and Cai, 2014). Secondly, essays that did not respond to the writing prompt were excluded. Thirdly, students who wrote an essay in Arabic and did not write in English were excluded. Lastly, students who spent more than 50 minutes per essay were excluded. Some texts were also submitted as hand-written pieces, manually typed by the team members. We created a balanced text set of 195 English and 195 Arabic essays.

The Arabic and English texts underwent the following annotations and amendments before analysis. Headings and titles were removed, as was any text that repeated the task instructions or the writing prompt. Secondly, we decided to correct spelling before analysis to ensure that we could capture our intended metadiscourse features. These corrections were made manually by reading through texts and systematically correcting spelling. Examples of spelling changes included words such as 'example' (misspellings included: 'exmple'), 'appear' (misspellings included: 'apper') and 'so' (misspellings included: 'sp'). Thirdly, spelling was standardised to American English as this was the common spelling used in most texts. However, no grammatical accuracy or turn of phrase accuracy was made. The raw and amended text is available to be downloaded from the Linguistic Data Consortium[1].

The QCAW corpus is also annotated for Part-of-Speech (POS) annotation given the interest from the community in having POS annotated Datasets. We used the QCRI Farasa POS annotation tool which is highly accurate (Darwish et al. 2017). The QCAW text format is encoded in UTF-8 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).TreeTagger was used to annotate the English texts, and Farasa was used to annotate the Arabic texts. Both tools are internationally recognised for accuracy and are widely used by researchers. Both raw texts and annotated texts are provided for all users. Details of the corpus are presented in Table 2 and 3. A breakdown of English and Arabic writing quality grades is shown in Table 4.

4. Main Findings

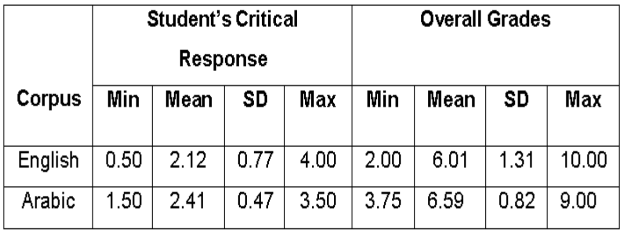
The Interpretation of learner corpus data differs from native language in many respects, among which the most important are form variability, learner errors, and writing assessment. In the case of learner English studies, form variability is usually investigated in terms of variations of linguistic forms or language-specific variations.

Corpus	Texts	Avg. 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

Table 2. English-Arabic corpus make-up

Corpus	Holistic Voice Score Breakdown			
	Min	Mean	SD	Max
ENG	3.00	3.48	0.53	5.00
ARAB	1.00	3.86	0.74	5.00

Table 3. Holistic voice score breakdown for English and Arabic corpora



Corpus	Student's Critical Response				Overall Grades			
	Min	Mean	SD	Max	Min	Mean	SD	Max
English	0.50	2.12	0.77	4.00	2.00	6.01	1.31	10.00
Arabic	1.50	2.41	0.47	3.50	3.75	6.59	0.82	9.00

Table 4. Grade breakdown for the English and Arabic corpora.

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 to investigate the longitudinal progress of L1 specific learners and to improve teaching strategies. To contribute to the aspects above, we will demonstrate how to use QCAW as a source of a learner corpus.

4.1 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 et al., 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; Biber, 2010; Gilquin & Granger, 2015; Gries & Wulff, 2020; Hendriks, 2005; Jarvis, 2000; Mollin, 2006; Paquot & Fairon, 2006; Pendar & Chapelle, 2008; Regan, 2013; Vyatkina, 2013; Wulff & 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] https://catalog.ldc.upenn.edu/LDC2022T04

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

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

(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.*

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. We further compared instances found in QCAW with argumentative writing by American university students provided by The Louvain Corpus of Native English Essays (LOCNESS)[2]. Unsurprisingly, we found that "on the first hand" was not used by native speakers.

4.2 Learner Errors and Learner English Corpus

Upon the analysis of our corpus, we observed that learner errors refer to uses of words/expressions that differ from native speakers and cause confusion.

In some cases, the QCAW examples shows incorrect use of "as well as" and "therefore/thus" by learners. For example: "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." We also noticed typical errors like "It's highly debates issue to determine whether or not email and telephone has made communication between people less personal. QCAW offers context to analyze learner errors systematically which could facilitate research in this area.

4.3 Learner Corpus Data and Writing Assessment

The QCAW enables a wide range of analyses and research studies into Arabic learner writing development and L1-L2 transfer effects. Here we highlight some specific examples of how this corpus can be leveraged in contrastive rhetoric, automated scoring, machine translation, and pedagogical contexts.

QCAW can provide valuable insights for writing assessment of L1 Arabic and L2 English texts. We explain that learner corpus data does not represent real-world language use, since texts are produced in controlled environments focused on meaning and language learning. Therefore, a corpus of argumentative essays written by Arabic L1 speakers learning English will exhibit unique features that can inform writing assessment. For example, our analysis of the QCAW reveals that students tend to produce much longer sentences in their L2 English essays compared to native speakers. This likely reflects influence from their L1 Arabic writing. Such corpus analysis can help us determine aspects like syntactic complexity and coherence when evaluating these students' English writing proficiency. Additionally, the metadata available in QCAW allows for studying the impact of variables like gender on writing features and performance in ways assessment previously may not have considered. Overall, the design and coverage of the QCAW positions it as a valuable data source to help us develop more authentic, reliable, and relevant writing assessments for Arabic learners of English in academic contexts.

5. QCAW Applications

In this section we will discuss the building learner corpora like QCAW which can have an impact on language teaching, learning, and research.

5.1 Teaching

One major application, we highlighted was that analyzing the errors and language usage patterns in learner corpora allows researchers to gain a more detailed understanding of the language acquisition process. For example, by examining the writing samples in QCAW, we were able to identify common problems that L1 Arabic students face when writing argumentative essays in both their native language and in L2 English. These insights could help us develop tailored teaching materials and strategies to address these learners' needs.

Another key application we noted was that learner corpora like QCAW enable the creation of more authentic and relevant language learning materials and assessments. By profiling the actual writing abilities and styles of Arabic learners of English, we could ensure that textbooks, assignments, and tests are appropriately calibrated to their proficiency levels and interests. Similarly, we explained how QCAW could be leveraged to evaluate the effectiveness of teaching approaches and materials already in use by comparing corpus samples to language targeted by instruction.

We also discussed how utilizing learner corpora forces consideration of their inherent strengths and weaknesses. To successfully integrate QCAW into language classrooms, teachers must provide focused feedback on common mistakes and support students in analyzing their own corpus-informed learner profiles. Other implications we covered included the potential for QCAW to reveal insights into learner lexical development, use of cohesive devices and discourse markers, interlanguage influences, and longitudinal writing development patterns. In

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

conclusion, we argued that constructing specialized learner corpora like the Qatari Corpus of Argumentative Writing has significant upside for enhancing language education and advancing understanding of student writing development in both L1 and L2.

The parallel Arabic-English essays allow direct investigation of rhetorical and discourse differences between the two languages. Researchers can analyze organizational patterns, cohesive devices, argumentation strategies, and metadiscourse markers in students' L1 vs L2 texts. For instance, Al-Jabr (2013) found Arabic writers tend to favor longer, more complex sentences than English writers. The QCAW provides robust data to test if such differences emerge in the Arabic and English essays on the same topics. Findings can enhance understanding of cross-cultural academic writing.

5.2 Automated Essay Scoring

The corpus can facilitate developing and testing automated scoring models for argumentative writing by Arabic learners. The rated essays and metadata represent valuable training and evaluation data for exploring different algorithms, feature sets, and prompt conditions. Models can predict holistic scores or analyze specific dimensions like coherence, lexical complexity, and grammar. Alikaniotis et al. (2016) demonstrated using learner corpora for this task. The QCAW specifically enables more valid AES for Qatari learner populations.

5.3 Machine Translation

The parallel Arabic-English corpus offers opportunities for improving machine translation quality for argumentative writing. The data can help adapt systems to better handle language-specific features, complex syntax, rhetorical conventions, and vocabulary in each language. Students' common errors flagged during translation (e.g. incorrect verb forms) can also inform automated feedback and grammatical error correction systems.

5.4 Writing Pedagogy

Analyses of the corpus can reveal Arabic writers' strengths, weaknesses, and developmental needs to guide curriculum and materials design. For instance, investigating text organization patterns or argument structure use can inform teaching of essay structure. Studying lexico-grammatical errors can help prioritize instruction in certain vocabulary, grammar rules, and language functions. The data allows assessment of learners' skills based on empirical evidence rather than assumptions alone. Findings can directly shape writing instruction and learning resources tailored for Qatari student populations.

In summary, the QCAW enables multifaceted lines of research into L1-L2 contrasts, learner language systems, automated writing evaluation, and pedagogical interventions tailored to Arabic learners' specific needs. Both the design and population coverage of this corpus facilitate investigations that can directly impact Arabic writing education and support learners in Qatar and beyond.

6. Limitations

Some limitations should be acknowledged regarding the QCAW corpus. First, the sample is heavily skewed toward female writers, reflecting the gender imbalance in the university population. Future efforts to balance gender representation could enhance the corpus. Second, the essays represent a snapshot of student writing rather than a longitudinal view of development over time. A corpus with multiple writing samples per student could enable more robust study of acquisition patterns. Finally, the rating measures for writing quality and voice, while systematic, depend on subjective human judgement. Developing automated or AI-assisted rating procedures could further strengthen the corpus annotations.

7. Conclusion

This paper has presented the development of the Qatari Corpus of Argumentative Writing (QCAW), a new bilingual learner corpus comprising 195 argumentative essays each in L1 Arabic and L2 English authored by Qatari university students. The corpus design, contents, and initial analyses offer several valuable contributions. First, the QCAW represents the first publicly available parallel corpus of Arabic and English argumentative writing with texts produced by the same learners in both languages. This enables more robust comparative research into Arabic-English contrastive rhetoric, transfer effects, and developmental patterns in acquiring writing proficiency in both languages. Second, the corpus compilation followed rigorous procedures for ethical collection of student texts, topic selection, rating for writing quality and voice, and formatting and annotation. The resulting resource provides a high-quality dataset to support diverse studies in automated writing evaluation, error analysis, discourse analysis, and more. Third, preliminary findings from the corpus offer insights into the distinctive features of Arabic learners' English writing, including form variability, common lexical and grammatical errors, and L1-specific influences. The corpus affords rich opportunities for deeper investigation of Arabic learner language using computational and empirical methods.Finally, the availability of detailed metadata opens up new possibilities for studying how learner factors like gender, academic major, and proficiency influence Arabic and English writing development. In conjunction with the essay texts and ratings, this enables more granular analyses of learner needs.

In summary, the QCAW learner corpus provides a valuable new resource for researchers and educators seeking to enhance the teaching and assessment of Arabic argumentative writing. The corpus design allows for multifaceted investigations into cross-linguistic writing development patterns, learner error profiles, and effects of individual variables. We hope that ongoing analysis of the QCAW will yield data-driven insights to inform writing pedagogy for Arabic learners and guide future corpus compilation efforts.

8. Data Availability

The dataset discussed in this paper (Ahmed et al., 2022), known as the Qatari Corpus of Argumentative Writing, is publicly accessible. This corpus was collaboratively developed by Qatar University, the University of Exeter, and Hamad Bin Khalifa University. It encompasses a diverse collection of argumentative essays in both Arabic and English, crafted by undergraduate students. These essays are accompanied by comprehensive annotations and essential metadata, providing insights into various aspects such as the students' linguistic backgrounds and the contextual details of the essays.

Interested researchers and practitioners can obtain this dataset from the Linguistic Data Consortium (LDC). For convenience and direct access, the dataset is cataloged under the identifier LDC2022T04. To explore or download the corpus, please visit the following URL: https://catalog.ldc.upenn.edu/LDC2022T04.This repository ensures that the data is preserved in a structured and standardized format, facilitating ease of access and utilization for academic and research purposes.

9. Ethical Considerations

In constructing and releasing the QCAW corpus, rigorous procedures were followed to protect student privacy and ensure the data was collected and is used ethically. Participation was completely voluntary, and informed written consent obtained from all students. Identifying information was removed from essay texts before analysis. All annotation and corpus access procedures were approved by the university IRB oversight board. Publicly releasing the corpus for research purposes was deemed to carry minimal risk, and students understood their writing may be included. However, researchers accessing the corpus should take care not to attempt to identify or contact individual students. Any excerpts reproduced in publications should be anonymized. With appropriate safeguards in place around data privacy and ethics, we believe the benefits of sharing this unique resource outweigh the risks.

10. Acknowledgement

This paper is the outcome of a funded research grant by the from Qatar National Research Fund (QNRF), now known as the Qatar Research Development and Innovation Council (QRDI) with the NPRP Grant: NPRP11S-1112-170006.

11. References

Abu Rabia, S., & Awwad, J. (2004). Morphological structures in visual word recognition: The case of Arabic. Journal of Research in Reading, 27(3), 321-336.

Ahmed, A. (2010). Contextual challenges to Egyptian students' writing development. International Journal of Arts and Sciences, 3(14), 503-522.

Ahmed, A. (2010). 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, UK.

Ahmed, A. M., Zhang, X., Rezk, L. M., & Pearson, W. S. (2023). Transition markers in Qatari university students' argumentative writing: A crosslinguistic analysis of L1 Arabic and L2 English. Ampersand, 10, 100110.

Ahmed, A., & Myhill, D. (2016). The impact of the socio-cultural context on L2 English writing of Egyptian university students. Learning, Culture and Social Interaction, 11, 117-129.

Ahmed, A., Debra Myhill, Esmaeel Abdollahzadeh, Lee McCallum, Wajdi Zaghouani, Lameya Rezk, Anissa Jrad, Xiao Zhang (2022), Qatari Corpus of Argumentative Writing, https://catalog.ldc.upenn.edu/LDC2022T04

Ahmed, A., Zhang, X., Rezk, L. and Zaghouani, W. (2023) Building an Annotated L1 Arabic/L2 English Bilingual Writer Corpus: The Qatari Corpus of Argumentative Writing (QCAW). Corpus-based Studies across Humanities, Vol. 1 (Issue 1), pp. 183-215. https://doi.org/10.1515/csh-2023-0012

Ahmed, A., Ali, N., Alzubaidi, M., Zaghouani, W., Abd-alrazaq, A. A., & Househ, M. (2022). Freely available Arabic corpora: A scoping review. Computer Methods and Programs in Biomedicine Update, 2, 100049. Elsevier.

Al-Mulla, S., & Zaghouani, W. (2020). Building a Corpus of Qatari Arabic Expressions. Proceedings of the LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020, 24.

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 & Curriculum. 13:2, 121-125.

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, UK.

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), 482-496.

Alqahtni, H. M. (2014). The structure and context of idiomatic expressions in the Saudi press (Doctoral dissertation, University of Leeds).

Caines, A., McCarthy, M., & O'Keeffe, A. (2016). Spoken language corpora and pedagogical applications. The Routledge Handbook of Language Learning and Technology. Abingdon: Routledge, 348-361.

Callies, M. (2013). Advancing the research agenda of interlanguage pragmatics: The role of learner corpora. Yearbook of corpus linguistics and pragmatics 2013: new domains and methodologies, 9-36.

Chen, A., & Gey, F. C. (2002, November). Building an Arabic Stemmer for Information Retrieval. In TREC (Vol. 2002, pp. 631-639).

Coxhead, A. (2000). A New Academic Word List. TESOL Quarterly, 34(2), 213-238.

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.

Dashtestani, R., & Stojkovic, N. (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), 435-456.

Darwish, Kareem, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki, "Arabic pos tagging: Don’t abandon feature engineering just yet," In Proceedings of the Third Arabic Natural Language Processing Workshop, 2017, (pp. 130-137).

Diab, M., Mansouri, A., Palmer, M., Babko-Malaya, O., Zaghouani, W., Bies, A., & Maamouri, M. (2008). A pilot arabic propbank. Proceedings of the 7th International Conference on Language Resources and Evaluation. Citeseer.

Maamouri, M., Zaghouani, W., Cavalli-Sforza, V., Graff, D., & Ciul, M. (2012). Developing ARET: an NLP-based educational tool set for Arabic reading enhancement. Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, 127-135.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., & Obeid, O. (2014). The first QALB shared task on automatic text correction for Arabic. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 39-47.

Gamon, M., Chodorow, M., Leacock, C., Tetreault, J., Ballier, N., Dıaz-Negrillo, A., & Thompson, P. (2013). Using learner corpora for automatic error detection and correction. Automatic treatment and analysis of learner corpus data, 127-150.

Gilquin, G. (2016). Discourse markers in L2 English. New approaches to English linguistics: Building bridges, 213-249.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. Journal of English for Academic Purposes, 6(4), 319-335.

Gilquin, G., & Granger, S. (2015). Learner language. In Biber, D. & Reppen, The Cambridge Handbook of English Corpus Linguistics (pp. 418–35). Cambridge University Press.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing EAP pedagogy. Journal of English for Academic Purposes, 6(4), 319-335.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising syner CALICO Journal, 465-480.

Granger, S. (2009). The contribution of learner corpora to second language a and foreign language teaching. Corpora and Language Teaching, 33, 13-32.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerised bilingual and learner corpora. In: K. Aijmer, B. Altenberg and M.

Johansson (Eds.), Languages in contrast. Papers from a symposium on text-based crosslinguistic studies (pp.37-51). Lund 4-5 March 1994. Lund UniversityPress.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (Ed.) Learner English on Computer (pp. 3-18).Addison Wesley Longman.

Granger, S. (2002). Learner English on Computer. London: Longman.

Granger, S. (2011). How to use foreign and second language learner corpora. Research methods in second language acquisition: A practical guide, 5-29.

Granger, S., Gilquin, G., & Fanny, M. (2015). Introduction: Learner corpus research –past, present and future. In: Granger, S., Gilquin, G., & Meunier, F. (Eds.), The Cambridge Handbook of Learner Corpus Research (pp.1–5). Cambridge University Press.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). The Cambridge Handbook of Learner Corpus Research. Cambridge University Press.

Gries, S. T., & Wulff, S. (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), 279–299.

Habash, N., Soudi, A., & Buckwalter, T. (2007). On Arabic transliteration. Arabic Computational Morphology: Knowledge-based and empirical methods, 15-22.

Habash, N., & Palfreyman, D. (2022). ZAEBUC: An Annotated Arabic-English Bilingual Writer Corpus. In N. Calzolari, F. Bechet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), 2022 Language Resources and Evaluation Conference, LREC 2022 (pp. 79-88). (2022 Language Resources and Evaluation Conference, LREC 2022). European Language Resources Association (ELRA).

Hendriks, H. (Ed.). (2005). The structure of learner varieties. Mouton-de Gruyter.

Hernández, P. S., & Paredes, P. F. P. (2005). Examining English for Academic purposes students' vocabulary output: corpus-aided analysis and learner corpora. Revista española de lingüística aplicada, (1), 201-212.

Hertel, T. J. (2003). Lexical and discourse factors in the second language acquisition of Spanish word order. Second Language Research, 19(4), 273-304.

Helmut Schmid (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Hyland, K. (1998). Hedging in academic writing and EAP textbooks. English for Specific Purposes, 17(1), 3-25.

Hyland, K. (2003). Second Language Writing. Cambridge University Press.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. Journal of Second Language Writing, 6(1), 183-205.

James, C. (1980). Contrastive analysis. 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., & Mehlmauer-Larcher, B. (2005). Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. ReCALL, 17(1), 65-84.

Kasper, G., & Rose, K. R. (2002). Pragmatics in Language Teaching. Cambridge University Press.

Kayaoglu, M. (2013). The use of corpus for close synonyms. Journal of Language and Linguistic Studies, 9(1).

Kaye, A. S. (2017). Arabic. In The world's major languages (pp. 576-593). Routledge.

Khorsheed, M. S. (2002). Off-line Arabic character recognition–a review. Pattern Analysis & Applications, 5, 31-45.

Khuwaileh, A. A., & Shoumali, A. A. (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), 174-183.

Kirmizi, O., & Karci, B. (2017). An investigation of Turkish higher education EFL learners' linguistic and lexical errors. Educational Process: International Journal, 6(4), 35.

Kuo, T. (2014). The acquisition of English past tense by Chinese EFL learners: A corpus-based study. International Journal of Corpus Linguistics, 19(3), 348-370.

Kusters, W.(2003).Linguistic Complexity. Netherlands Graduate School of Linguistics.

Laufer, B. (1997). The lexical plight in second language reading. In J. Coady & T.

Huckin (Eds.), Second language vocabulary acquisition: A rationale for pedagogy (pp. 20-34). Cambridge University Press.

Lefebvre, C., & Al-Tonsi, A. (2008). Sociolinguistics and language education in the Arab world. Bristol: Multilingual Matters.

Lennon, P. (1991). Error: Some problems of definition, identification, and distinction. Applied Linguistics, 12(2), 180–196.

Li, S. (2017). Using corpora to develop learners' collocational competence. Language Learning & Technology, 21(3), 153–171.

Man, D., & Chau, M. H. (2019). Learning to evaluate through that-clauses: Evidence from a longitudinal learner corpus. Journal of English for Academic Purposes, 37, 22-33.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyse language use. Annual Review of Applied Linguistics, 39, 74–92.

Meunier, F. (2015). Developmental patterns in learner corpora. In: Granger, S., Gilquin, G., & Meunier, F. (Eds.), The Cambridge Handbook of Learner Corpus Research (pp. 379–400). Cambridge University Press.

Mollin, S. (2006), Euro-English: Assessing variety status. Tübingen: Gunter NarrVerlag.

Mourssi, A. (2013). Crosslinguistic Influence 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 & Practice in Language Studies, 3(3).

Müller-Hartmann, A., & Schocker-von Ditfurth, M. (2011). Introduction to English language teaching: Optimise your exam preparation. Stuttgart: Klett Lerntraining.

Muysken, P. (2000). Bilingual speech: A typology of code-mixing. Cambridge University Press.

Nation, I. S. (2001). Learning vocabulary in another language. Cambridge University Press.

Naz, S., Umar, A. I., Shirazi, S. H., Ahmed, S. B., Razzak, M. I., & Siddiqi, I. (2016). Segmentation techniques for recognition of Arabic-like scripts: A comprehensive survey. Education and Information Technologies, 21, 1225-1241.

Obeid, O., Zaghouani, W., Mohit, B., Habash, N., Oflazer, K., & Tomeh, N. (2013). A Web-based Annotation Framework for Large-scale Text Correction. Sixth International Joint Conference on Natural Language Processing, 1.

Okamoto, K. (2010, July). Incorporating corpora into English language teaching for undergraduate computer science and engineering students with limited proficiency. In 2010 IEEE International Professional Communication Conference (pp. 152-156). IEEE.

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.

Palmer, M., Babko-Malaya, O., Bies, A., Diab, M. T., Maamouri, M., Mansouri, A., & Zaghouani, W. (2008). A Pilot Arabic Propbank. LREC.

Paquot, M. & Fairon, C. (2006). Investigating L1-induced learner variability: Using the web as a source of L1 comparable data. 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 crosslinguistic perspective. In F. Meunier and S. Granger (Eds.), Phraseology in foreign language learning and teaching (pp. 101-119). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Pendar, N. & Chapelle, C. A. (2008). Investigating the promise of learner corpora: Methodological issues. CALICO Journal, 25(2), 189–206.

Phoocharoensil, S. (2013). Crosslinguistic Influence: Its Impact on L2 English Collocation Production. English Language Teaching, 6(1), 1-10.

Rangel, F., Rosso, P., Zaghouani, W., & Charfi, A. (2020). Fine-grained analysis of language varieties and demographics. Natural Language Engineering, 26(6), 641-661. Cambridge University Press.

Regan, V. (2013). Variation. In Herschensohn, J. & Young-Scholten, M. (Eds.), The Cambridge Handbook of Second Language Acquisition (pp.272–91). Cambridge University Press.

Römer, U. (2010). The acquisition of English articles by German learners: A corpus-based study. International Journal of Corpus Linguistics, 15(3), 386-408.

Rozovskaya, A., Bouamor, H., Habash, N., Zaghouani, W., & Obeid, O. (2015). The second QALB shared task on automatic text correction for Arabic. Proceedings of the Second workshop on Arabic natural language processing, 26-35.

Saiegh-Haddad, E., & Henkin-Roitfarb, R. (2014). The structure of Arabic language and orthography. Handbook of Arabic literacy: Insights and perspectives, 3-28.

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., & Atwell, E. (2013). A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging. Word Structure, 6(1), 43-99.

Schmitt, N. (2000). Vocabulary in language teaching. Cambridge University Press.

Shirato, J., & Stapleton, P. (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 University Press.

Smirnova, E. A. (2017). Using corpora in EFL classrooms: The case study of IELTS preparation. RELC Journal, 48(3), 302-310.

Sorace, A. (2011). Pinning down the concept of overgeneralisation. Studies in Second Language Acquisition, 33(2), 231-262.

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

Valero Garcés, C. (1997). The interlanguage of Spanish students beginning English Philology. GRETA, 5(2), 74-78.

Vaughan, E., & Clancy, B. (2013). Small corpora and pragmatics. Yearbook of corpus linguistics and pragmatics 2013: New domains and methodologies, 53-73.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. The Modern Language Journal, 97(S1), 11–30.

Wei, L. (2018). Contrastive learner corpus research: A review. Language Teaching, 51(4), 446-471.

Wulff, S., & Gries, S. (2021). Exploring individual variation in learner corpus research: Methodological suggestions. In B. Le Bruyn & M. Paquot (Eds.), Learner corpus research meets second language acquisition (pp. 191-213). Cambridge University Press.

Yoo, I. W., & Shin, Y. K. (2019). Determiner Use in English Quantificational Expressions: A Corpus-Based Study. TESOL Quarterly, 54: 90-117.

Yoon, H., & Jo, J. W. (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. Language Learning & 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)

Zaghouani, W. (2014). Critical survey of the freely available Arabic corpora. International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Reykjavik, Iceland, 26-31 May 2014.

Zaghouani, W., & Dukes, K. (2014). Can Crowdsourcing be used for Effective Annotation of Arabic? LREC.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., & Palmer, M. (2010). The revised arabic propbank. Proceedings of the fourth linguistic annotation workshop, 222-226.

Zaghouani, W., Hawwari, A., & Diab, M. (2012). A pilot propbank annotation for quranic arabic. Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature, 78-83.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., & Oflazer, K. (2014). Large Scale Arabic Error Annotation: Guidelines and Framework.

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., & Oflazer, K. (2015). Correction annotation for non-native arabic texts: Guidelines and corpus. Proceedings of The 9th Linguistic Annotation Workshop, 129-139.