UDC 81`33 DOI: 10.18413/2313-8912-2024-10-3-0-6

Abdulaziz B Sanosi[1]
Abuelgasim Sabah Elsaid Mohammed[2]

ASAWEC: towards a corpus of Arab scholars' academic written English

Prince Sattam bin Abdulaziz University,
5229, Hawtat Bani Tamim, 16628, Saudi Arabia
E-mail: a.assanosi@psau.edu.sa
ORCID:0000-0003-3447-2818

Prince Sattam bin Abdulaziz University,
7524 Hawtat Bani Tamim, 16628, Saudi Arabia
E-mail: a.ibrahim@psau.edu.sa
ORCID: 0000-0001-9791-0905

Received 07 October 2023; accepted 15 September 2024; published 30 September 2024

Abstract: Linguistic corpora have been used in a wide range in recent years. Different types of linguistics analyses in both spoken and written discourses are being conducted using the corpus linguistics approach. Among these, academic writing has received considerable attention. Corpus linguistics has provided insights into the academic writing of both native and non-native English language learners and writers in general. Nevertheless, relatively few studies have investigated this topic in the Arab EFL setting. Consequently, there is a relative paucity in corpora of Academic written English by Arab speakers. To address this gap, we compiled the Arab Scholars' Academic Written English Corpus (ASAWEC) which is a specialized corpus of Arab scholars' academic written English. We collected the corpus texts according to specific criteria, and then we normalized and cleaned the data. The texts were then tokenized and tagged and the corpus underwent initial tests which yields insightful findings on Arab scholars' academic written English such as the low lexical diversity and the utilization of various discourse techniques. The present paper introduces the corpus, provides details on its compilation, presents initial results and statistics, and discusses potential limitations and future perspectives for updating the corpus. It is envisaged that this project will encourage the use of the ASAWEC and help in launching similar initiatives to advance research in Arab corpus linguistics.

Keywords: Specialized corpus; Corpus compiling; Academic writing; Corpus linguistics; L2 writing

How to cite: Sanosi, A. B, Mohammed, A. S. E. (2024). ASAWEC: towards a corpus of Arab scholars' academic written English, *Research Result. Theoretical and Applied Linguistics*, 10 (3), 116-134. DOI: 10.18413/2313-8912-2024-10-3-0-6

Introduction

The shift from formal linguistics to functional linguistics paved the way for new methods of analyzing language, the most prominent of which is corpus linguistics. Functional linguistics emphasizes the study of language as a functional system through empirical evidence that relies on real contexts. Accordingly, the need for actual language data among linguists was a significant driving force in the development of modern corpora by the 1950s (Blecha, 2012). Corpus linguistics helps in meeting this need as it is concerned with "the study of language based on examples of "real life" language use" (McEnery and Wilson 2001: 1) using large, structured corpora that reflect actual language use and show patterns of language employed by different users.

Currently, linguists from various schools rely on linguistic corpora not only to gather authentic language examples but also to test their linguistic theories against quantitative data derived from actual language usage (Kubler and Zinsmeister, 2015). The huge advances in computers and related technologies make this possible and enhance the practice. Corpora have now become of many types and large sizes amounting to tens of millions and more. Nevertheless, the reality of corpus linguistic analysis in the Arab world is still humble as few corpora of different types are available in the Arab EFL setting (Almohizea, 2017; Alotaibi, 2017). The researchers think this is unfortunate, especially after the massive development in corpus software and Natural Language Processing (NLP) solutions that made linguistic data collection and analysis a relatively convenient process (Dunn, 2022).

Thus, the problem of this research can be formulated as corpus linguistic researchers in the Arab EFL lack access to robust and well-established corpora that can be relied upon for linguistic analysis. This problem may dissuade scholars from pursuing corpus-based studies because their only alternative would be to create their own DIY corpora, which is not a feasible option for many researchers as it is a time and effort-consuming task. Even for those who opt to develop their corpora, this approach may lead to unsatisfactory results, as compiling a corpus involves numerous intricate steps to ensure its size appropriateness, representativeness, balance, and inclusivity. Moreover, the issue is intensified by the fact that corpus linguistics requires expertise in other disciplines, particularly in IT and related fields. Even though many corpus linguists and applied linguistics researchers excel in their respective areas, a deficiency in computer science and programming expertise may hinder their ability to develop well-prepared corpora. This is because new corpora heavily depend on modern technology, which is constantly evolving and being updated. Adapting to these changes poses an additional challenge for linguistic researchers.

Furthermore, to cope with the trending approaches of linguistic analysis and corpus-based discourse studies that guide most of today's applied linguistic research, further interest is needed in the field of corpus linguistics in the Arab EFL setting. This interest entails compiling different types of corpora as "[the] number of readily available specialized corpora is still much smaller than [the] number of general corpora" (Blecha, 2012: 5), and even the existing ones cannot meet all the scholars' research needs. Thus, the primary aim of this paper is to present the specialized Arab Scholar Academic Written English Corpus (ASAWEC) to the public and describe the processes and procedures taken to build it. The project is anticipated to lay the foundation for a large, specialized corpus of academic writing by Arab scholars and inspire similar initiatives in the field. Additionally, the main output of the project, ASAWEC, will be accessible to the public to promote research in the area and allow for the corpus to be revised and enhanced.

Literature Review

The word *corpus* is a Latin word that means *body*. However, in modern linguistics, a corpus is not a mere body of texts, it rather involves four principles which are: sampling and representativeness, being of a finite size, machine-readable, and having a standard reference (McEnery and Wilson, 2001). These features enable linguistic corpus to provide an evidence-based source of linguistic analysis (Meyer, 2023), and hence make corpus linguistic "scientific method of language analysis [that] requires the analyst to provide empirical evidence in the form of data drawn from language corpora in support of any statement made about language" (Brezina. , 2018: 2). The involvement of the scientific method and specific properties of linguistic corpus entails a new contemporary definition that accounts for all these variables and keeps it apart from the traditional views towards corpora and corpus linguistic analysis which entail many descriptions. This definition states that "Corpus linguistics is the investigation of linguistic research questions that have been framed in terms of the conditional distribution of linguistic phenomena in a linguistic corpus." (Stefanowitsch, 2020: 56).

As corpus linguistics has expanded in recent decades to incorporate almost all types of linguistic analysis, various kinds of corpora have been classified and standardized by researchers in terms of their register or purpose.

Types of Corpora

A clear-cut classification of corpora depends on their register i.e., spoken or written. Most available corpora are, due to practical reasons, written. Sinclaire (1991) stated that compiling a spoken corpus is not as a straightforward task as a written one since it involves many challenges such as transcription and representativeness. He argued that collecting actual spoken language is impossible, whereas collecting quasi-speech material, such as film transcripts, is considered artificial and does not represent actual language use. Although there is a significant advance in transcription, speech recognition, and speech-to-text applications, still spoken corpora are rare compared to written (Lemmenmeier-Batinić, 2023).

According to Myer's (2023) classification, there are several types of commonly available English corpora. Multipurpose corpora are designed to represent a particular register of speech or writing and can be used for various analyses. Learner corpora contain written and/or spoken language produced by learners of the language, also known as "interlanguage" (Gilquin, 2020). Teachers, researchers, and material designers use learner corpora to inform their teaching practice, gather data on students' language patterns, and enhance their syllabi. The main focus of corpus-based instruction is writing (Liu, 2022), and teachers use learner corpora as a teaching strategy to help students analyze concordance lines and find patterns of language use. Alternatively, they may use corpus-based teaching materials to improve students' academic writing skills.

Historical corpora are collections of texts from a specific time that can be compared to other collections to study linguistic evolution over time. Parallel corpora contain texts in English and their translations into another language. They are used to enhance translation skills and study variations across languages.

Another widely recognized type of corpora is a specialized corpus (Baker, et al., 2006; Blecha, 2012; Stefanowitsch, 2020) which is more relevant to the current study.

Specialized Corpora

In its essence, a specialized corpus is "a corpus which has been designed for a particular research project, for example, lexicography for dictionary compilation, or to study particular specialist genres of language" (Baker, et al., 2006: 147). Although this type does not contribute to the description of language as a whole since they do not represent the usual natural language use, it is nevertheless effective in describing features of the language that distinguish specific genres or linguistic communities (Blecha, 2012). These features place restrictions on the texts to be included within them and hence they are normally smaller than general or reference corpora (Baker, 2010).

Specialized corpora aim to identify domain-specific vocabulary, as researchers and educators have observed that words vary across different fields. By creating wordlists from various specialized corpora, it is possible to narrow down the meanings of words associated with each context. In this regard, several researchers compiled corpora addressing specific fields and the language used in them. Toriida (2016) described the steps she followed in compiling a specialized corpus for the EAP nursing program. She found that the corpus was effective in generating an annotated wordlist for academic English in the field of nursing.

Specialized corpora can also have practical purposes, as demonstrated by Allan et al. (2023), who reported the construction of a specialized corpus of written English from Swedish year 9 National Tests. Although the corpus appears to be a learner corpus, its focus on test language and inclusion of teacher feedback suggests that it can also be used as a tool to guide teachers in assessing tests which can be another motive for compiling specialized corpora.

Specialized corpora and academic writing

The use of corpus linguistics in the investigation of academic writing is a common practice. This is mainly because this type of writing necessitates strict criteria to differentiate it from other forms of writing. Specifically, it should demonstrate proficiency in rhetorical conventions, linguistic characteristics, vocabulary, and syntax (Utkina, 2021). Since such aspects are unusual in common everyday language, they represent an ideal source for specialized corpora. Several studies explored the use of specialized corpora in academic writing. For example, Guerra and Smirnova (2023) compiled a 1.5-million-word corpus to analyze the linguistic complexity in professional academic writing. They compared the research articles published in peer-reviewed journals in hard and soft disciplines. They found the specialized corpus was informative and presented empirical evidence on the linguistic variation.

Other specialized corpora merged the two aims of analyzing academic writing and generating wordlists. In this regard, Jamalzadeh and Tabrizi (2020) created a specialized corpus of 3.7 million words with the goal of developing a Tourism Academic Word List (TAWL) that includes the most used academic vocabulary in various sub-disciplines of tourism. They achieved this by analyzing a written corpus of academic research articles in the field of tourism.

Considering research articles written by non-native speakers of English, the issue seems more interesting as it is anticipated that specific features of such forms of writing may be showcased. Research has not neglected this issue, though very rare studies are available in this regard. For example, Fuentes (2009) constructed a specialized corpus to examine the academic writing of Spanish authors in the field of computer science. The study involved comparing the findings with a selection from the British National Corpus (BNC). The results indicated that there were no significant differences in academic vocabulary or style between native and non-native writers. However, it is important to note that the small size of the corpus (only 25,000 words) may raise concerns about the validity of these claims. Nevertheless, the study employed a rigorous data collection method by gathering research articles in their final version, prior to peer review. This approach enhances the authenticity of the data as peer review and proofreading can significantly impact the work of scholars.

In the Arab context, most studies that investigated Arab academic writing utilized learner corpora. There are few studies that investigated the use of Arab scholars of different linguistic features in their academic writing by compiling specialized research-specific corpora. Among these studies are Akeel (2020) who investigated the use of modal verbs by Saudi postgraduate writers of English. and Sanosi (2022) who investigated the use of lexical bundles by both Arab learners and British Scholars. Both authors compiled their own corpora for the research purposes and compared the results against the British Academic Written English (BAWE) corpus. Akeel (2020) found that Arab postgraduate writers of academic English underused modal verbs in their academic writing, while Sanosi (2022) found no effect of further study or professional practice on the development of lexical bundle use by Arab Scholars.

It can be concluded that specialized corpora can have a notable impact on improving academic writing for both non-native English students and scholars. As a result, this study draws inspiration from the insights discussed above regarding academic writing and specialized corpora. Thus, the rationale for the current study was formulated.

Study rationale

The initial motivation for compiling a specialized corpus of academic writing by Arab scholars was a research project on metadiscourse resources used by Arabic academic writers (in press). During the preparation of this research, it was discovered that data pertaining to Arab scholars' English writing was scarce despite their vast publication in recent years. This drove the researchers to compile a corpus for research purposes and make it available to their surrounding research community, as well as publicize it online.

Then, the rationale for compiling ASAWEC is threefold. First, the corpus is intended to be a valuable resource for researchers who aim to explore linguistic and discourse aspects of academic writing by Arab scholars. For example, analysis of the corpus may reveal common areas of difficulty for these writers, such as difficulties with certain grammatical structures or academic vocabulary. Additionally, the corpus can shed light on the linguistic resources that are recurrently used by Arab scholars writing in English, such as certain rhetorical strategies or discourse markers. Finally, analysis of the corpus can reveal the linguistic devices that Arab scholars use to mark stance and construct metadiscourse, which can provide insights into their rhetorical and communicative goals. This will enrich the field and answer research questions on the effects of cultural background on writing practice and discourse structure.

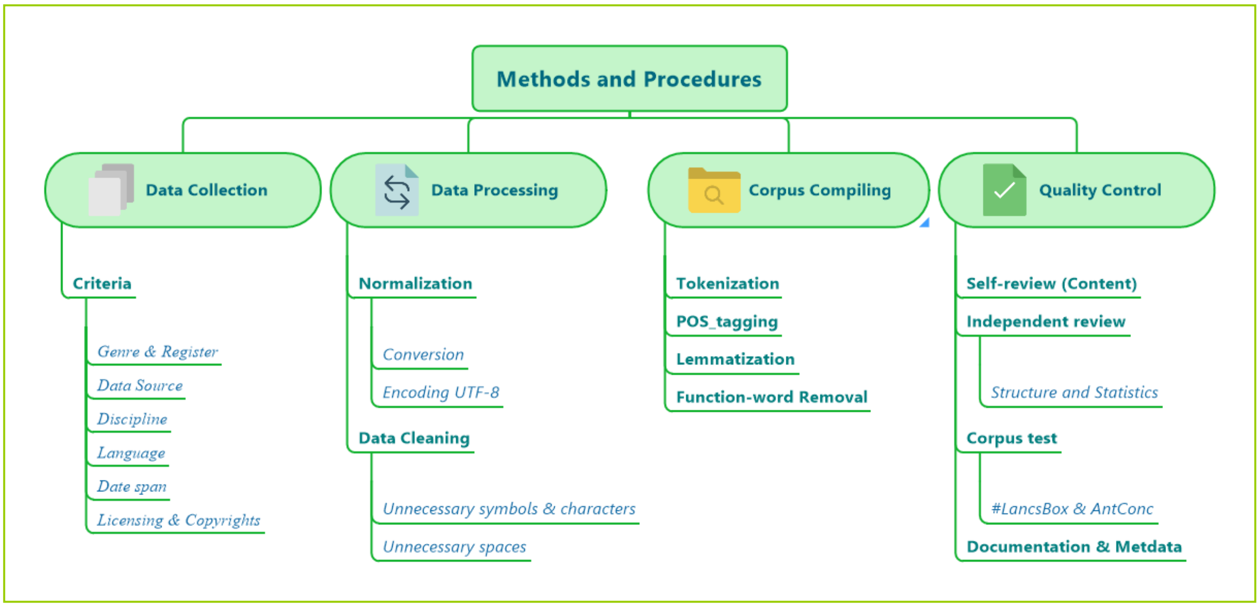
Second, the corpus can inform the development of training and teaching materials designed to enhance academic writing at college level and professional development intended for scholars. By identifying areas of difficulty for academic writers, curriculum designers and reviewers can update curricula and courses for both graduate and postgraduate students which can result in improved outcomes for scholars and students.

Finally, the corpus can contribute to promoting inclusivity in academic discourse by introducing a sample of the academic discourse of a group rarely represented in corpus studies. This aim will promote equitability in the academic community and promote diverse scholarly work.

Methodology

The researchers utilized the corpus compiling methods and techniques recommended by established scholars and pioneers in corpus linguistics, such as Sinclaire (1991), Atkins et al., (1992), McEnery and Wilson (2001), Baker et al. (2006), and Baker (2010). Studies by contemporary scholars in this field were also reviewed, e.g. Toriida (2016), Brezina (2018), Stefanowitsch (2020), Meyer (2023) and others. We also referenced resources on Python and NLTK, such as Bird et al. (2009) and Dunn (2022). Additionally, the researchers sought continuous input from colleagues in the IT and statistics fields to ensure the utmost accuracy and rigor in constructing ASAWEC, which involved utilizing various software and applications. Ultimately, the corpus was compiled following the procedures outlined in Figure 1.

Figure 1. Procedures of ASAWEC compiling



Thus, the methods for compiling the corpus and subsequent design of this research article are summarised in the steps outlined in Figure 1 and detailed below.

Sampling Frame and Data Collection:

The main step in the method is to establish a sampling criterion to guide the selection of the texts to be included in the corpus. Sampling of a corpus is one of the issues that face corpus compilers because regular statistical measures for defining sampling and population are hardly applicable to corpus linguistics mainly because it is difficult to define the volume of the whole language production that the sample is to be taken from (Brezina, 2018). However, in specialized corpora, the problem is less thorny, as it is noted that “the more highly specialized the language to be sampled in the corpus, the fewer will be the problems in defining the texts to be sampled” (Atkins et al., 1992: 7). To limit the language production that the corpus is to be taken from as a representative sample, researchers suggested using a sampling frame prior to the selection process. The sampling frame involves establishing a set of criteria to control the texts to be selected for building the corpus (Meyer, 2023). This practice is especially common in compiling specialized corpora since they need parameters to set limit to the text they would like to include in their corpora (Hunston, 2002). Accordingly, several criteria were adopted by the researchers to establish the sampling frame and control the data collection for the present project. These criteria are discussed below.

Genre and register

The aim of this project is to provide valid and accurate data to analyze the patterns and trends of the English writing styles used by scholars and academics whose L1 is Arabic. The motivating research topic was to analyze metadiscourse resources used by Arab scholars. However, it was envisaged that the corpus designed for the study could be a valuable source for future research in the field rather than a byproduct of one research. Further, the corpus can be a nucleus for a larger corpus after continuous updating and wider participation from potential partners. Accordingly, this aim informed the genre and register of the corpus and hence set the primary criterion for the sampling frame.

It was envisaged that the most practical and common representative of academic style is the scientific article genre, mostly in IMRaD style, that is published in online journals. Scholars who publish in such journals normally follow strict writing conventions and adhere to the formal use of vocabulary and grammar. These register-specific conventions are envisaged to stimulate several research questions such as investigating scholars' use of vocabulary, syntactic structures, discourse markers, collocations, lexical bundles, and stance markers to name a few. Exploring this will be more feasible by following a corpus-assisted discourse analysis approach and this is the objective that the current project aims to help in achieving. Plans to update the corpus include the addition of other genres such as conference papers, presentations, seminars and research notes. While these genres do not fully represent academic writing, it may not be suitable to include other genres such as essays, theses, and dissertations at this point, as they may reflect the writing of students rather than scholars.

Data source

The major source of the corpus is the scientific journals that were actively published in the region for the specified period. We targeted journals that are published by Arab universities and academic entities as they are more likely to incorporate research articles written by Arab scholars. Through the building and cleansing processes, we adjusted the number of selected articles to reach our initial target of one million words. The decision to initially select only Arab journals was based on practical considerations as it was more convenient to determine the identity of the writers. However, future updates will involve adding articles written by Arab scholars in international journals. This task will require further work, as it involves verifying the identity of the scholars, either through direct contact or their respective institutions. This process will be more feasible once institutional participation in the development of this corpus is achieved. This is one of the future plans of the corpus compilers.

Discipline

As a starting point we selected disciplines that related to the English language and literature. In this regard, we searched for journals whose scope is one of applied linguistics, literature, and translation. We believe that these will form a suitable foundation for the corpus as scholars in this field are more likely to be aware of academic writing conventions and their articles would serve our initial purpose of forming the corpus. However, the proposed updates, as outlined below, involve incorporating articles from the fields of science, social sciences, applied sciences, and arts. In this stage, we build a pool containing 19 journals.

Language and content

As explained before, the corpus compiling aimed to provide data for corpus linguists to analyse academic English written by Arab scholars. Therefore, we searched for publication containers that address linguistic issues in the Arab World in general. Consequently, most of these articles contained an amount of Arabic text that must be removed to maintain the corpus accuracy and representativeness. In some cases, the Arabic content is small and can be managed by the cleansing procedures e.g., Arabic abstracts, Arabic running heads, and Arabic headers and footers. However, other articles contain a huge amount of Arabic text that cannot be deleted without affecting the article's structure and meaning. This feature was found in two journals that focus on Quranic texts and translation. Both journals were excluded to maintain consistency and accuracy reducing the pool to 17 journals.

Date span

To provide more updated data we specified the period of publication to be between 2019 – 2023. We settled for this date span after continuous adjustments stemming from observing the publication of the journal selected in the first stage and the application of the other criteria. The time span for the articles is considered for the whole corpus since there are some journals which started publication in 2021 while there are others which discontinued publication before 2023. The implementation of this criterion reduced the number of journals in the selection to 11 journals.

Licensing and copyrights

Recently, many journals adopted an open access policy and used Creative Common (CC) licenses that facilitate the use of their content for compiling corpora. Some journals make their content available under a CC BY license which allows others to share, copy, redistribute, remix, transform, and build upon the original work for any purpose, even commercially, if they give appropriate credit to the original creator(s) and indicate if changes were made. Others adopt CC BY-NC, or CC BY-SA licenses, which enable the reader to browse, download, and reuse their contents as the purpose is non-commercial and the new work generated from them is shared under the same licenses. Yet, there are less permissive licenses such as CC BY-ND, and CC BY-NC-ND which are unsuitable for corpus compiling as they prohibit modification of any means to the original content unless after acquiring permission from the publishers.

Most journals in our pool are published under CC BY and CC BY-NC so we selected them for the study. For other restricted journals we contacted five publishers to request full access to their archives and we explained the necessary modifications that should be done in the cleansing stage to fit the files for the corpus building. We ask this after we explained the aims and details of our project confirming that it is strictly non-commercial in nature, and its primary purpose is to advance scholarly knowledge and understanding. According to these procedures, we got the approval of two journals. We also excluded three other journals, one of which declined our request while we got no responses from the other two. The pool of the study was reduced to 10 journals by this stage. Ultimately, we had 8 journals that met the criteria to build the nucleus of ASAWEC. The overall details of the chosen journals are organized in Table 1.

Table 1. Source Journals of ASAWEC

No.	Journal	Research Scope	Publisher
1	Algerian Translation and Languages Journal (ALTRALANG)	Language, Linguistics, and Translation	University of Oran 2 Mohamed Ben Ahmed, Algeria
2	Arab World English Language Journal (AWEJ)	English Language and Linguistics	Arab Society of English Language Studies (ASELS)
3	The Egyptian Journal of Linguistics & Translation (EJELT)	Linguistics, literature, and translation	Sohaj University, Egypt
4	International Journal of English Language & Translation Studies (IJELTS)	ELT, literature, linguistics, translation	University of Sebha, Libya.
5	Journal of English Studies in Arabia Felix (JESAF)	English language, Linguistics, and Literature	Arabia Felix Academy (ARAFA), Yemen
6	Journal of Research in Language and Translation (JRLT)	Linguistics, Translation, second language teaching & learning	King Saud University, Saudi Arabia
7	The Saudi Journal of Language Studies (SJLS)	Second language teaching and learning, ESP, linguistics	King Khaled University, Saudi Arabia
8	Journal of Umm Al-Qura University for Language Sciences and Literature (UQUJLL)	writing, verification, translation	Um Al-Qura University, Saudi Arabia

We set further criteria for article selection to ensure the quality and representativeness of the corpus. The first criterion was to take all possible measures to ensure that the writer of the article was an Arab scholar. While there were clear features that indicate a high possibility of this criterion, such as full name, bio, and affiliation, it was sometimes difficult to decide whether the writer was Arabian. The main reason for this is that Arab nationality does not guarantee an Arabic language background, as the Arab World population is ethnographic and linguistically complex. There are many countries where there are other languages and dialects used besides Arabic which may make the writer's L1 not Arabic. We took all these measures into account, and we selected only files that have enough indicators that they were written by a scholar whose first language is Arabic and excluded all files that are surely or likely written by non-Arab scholars. We use article metadata, research topics, and, in a few cases, personal correspondence, with some writers to guarantee the maximum possible representation.

All the files were downloaded to a local folder named ASAWEC PDF. The files were indexed alphabetically and sequentially. The file names are composed of the journal name abbreviation as a prefix followed by an underscore and a sequence number. Then the files were processed to compile the corpus.

Data Processing:

To prepare the files for the final corpus compilation, we utilized several processes and a variety of software and programming libraries. Each will be explained below in relation to the specific process in which it was used. While some of these processes were meant to build the corpus in the first place, other advanced processes were executed to provide preliminary results to test the corpus and generate statistical information for documentation purposes. Also, these advanced procedures were meant to provide different versions of the corpus to make it fit for various corpus linguistic analyses. These procedures are detailed below.

Normalization

The first step was to normalize the file format and encoding. First, all the PDF files were converted to text (.txt) format to ensure consistency and feasibility of analysis as this format is compatible with almost all common corpus linguistic software. This feature was performed using AntFileConverter (Anthony, 2022) which is a software application that converts batch PDF files to plain text files with the UTF-8 coding format. While this step made the files consistent in terms of both format and encoding, it nevertheless left files with many formatting characters that may negatively impact the automatic analysis of the corpus.

Data cleansing:

The data was cleaned both manually and automatically. Manual cleaning was standardized by removing the parts of the articles that are repeated or which have no value to the linguistic content, and they do not represent the writing of the authors. Accordingly, we deleted references, acknowledgements, appendices, and conflict of interest statements from each article. Also, articles bibliographical information, headers and footers including article titles, running heads, and journal information were deleted. All information, however, was retained in the metadata file which provides detailed information about each article.

The text files were subsequently cleansed using the Python 're' module, which is a built-in library that employs regular expression (RegEx) commands. The functions within this module enable users to locate all instances of particular text patterns and substitute them with designated text or white space, thereby facilitating a comprehensive cleaning process across the corpus. The cleaning processes include the following procedures:

1. Removing Arabic characters using a regex pattern.
2. Removing URLs.

3. Removing unnecessary white spaces by splitting the text into words and joining them back together with single spaces.

4. Removing unnecessary symbols and codes using a regex pattern.

5. Removing any remaining non-alphanumeric characters and noise data.

Afterwards, all the text files were stored in the corpus folder, and the subsequent steps of compiling the corpus were conducted.

Corpus compiling

The processed files were assembled and underwent basic corpus processes to compile and test ASAWEC. These processes produced the foundational results and statistics, which will be examined in the next section. Additionally, they led to the creation of multiple versions of ASAWEC, with the main folder being uploaded and made public, and other versions available upon request. The processes will be thoroughly discussed in the following subsections.

Tokenization

To create a standardized and consistent representation of the text files that is suitable for corpus linguistic analyses, the files were tokenized using Python and NLTK libraries. This library allows users to use several text processing functions. For this stage, we used the *word tokenize()* function to divide the texts into words or (tokens). This process was done as the previous processes of converting and cleaning may cause distortion in the distribution of words, and this can be solved by segmenting the units into tokens (Baker et al., 2006). The process was performed to break up text into isolated tokens (words) as most corpus linguistic analyses are carried out on the word level. However, punctuation marks were retained as the meant corpus is for academic writing where punctuation marks can provide valuable information on writing styles and text structures. The tokenized corpus was saved in an independent folder named ASAWEC_tokenized which is used to generate further copies of the ASAWEC to provide a wide range of options for future researchers to use the corpus.

Annotation

There are several types of linguistics annotation that are conducted to corpora on various levels ranging from morphemes to phrases or sentences. Researchers recognized that “the most common type of linguistic annotation is part-of-speech annotation which can be applied automatically with high accuracy, and which is already very helpful for linguistic analysis despite its shallow character” (Kubler and Zinsmeister, 2015: 22). Part of speech (POS) tagging is a process of assigning a word to its appropriate word class to aid in the analysis of a text (McEnery and Wilson, 2001). To make ASAWEC more convenient for future research, it was tagged using Python and NLTK libraries. The source files were taken from the resultant tokenized cleaned corpus to pave the way for the next steps which also aim to provide another version of the corpus that is suitable for a specific type of analysis. The process was achieved using the *pos_tag()* function from the NLTK library.

Lemmatization

Lemmatisation is a process of collating and counting the inflected and derived forms of words (Collins, 2019). It is considered vital to corpus linguistics because it facilitates the process of search and makes it more tractable and straightforward (Cox and Newman, 2020). As this corpus is intended for a wide range of research most of which requires searching at word level, it was deemed necessary to provide a lemmatized version of the corpus from its initial compiling steps. Accordingly, the tokenized and pos-tagged corpus was loaded to Python and lemmatized using the NLTK WordNetLemmatizer() function. The output files were stored in the third folder named ASAWEC_lemmas. The NLTK library is efficient in lemmatising since it provides one stance of each lemma across the corpus, so there is no need to lemmatize each file independently. Therefore, the generated corpus will be economical and convenient for research purposes.

Function words removal

When searching for frequency to determine keywords or to generate wordlists, the most frequent words that an analyst encounters are the function words such as *the*, *a*, and *of*. (Crawford and Csomay, 2016) especially in academic writing which is dominated by the nominal style that includes numerous articles and prepositions (Stefanowitsch, 2020). Although these words are highly frequent in English, their contribution to the unique meaning is low; therefore, they may provide misleading pictures in specific types of analysis e.g. semantic and sentiment analysis. Accordingly, it is a common tradition in corpus linguistic analysis to remove function words or stopwords as referred to in NLP literature (Dunn, 2022), when conducting types of analysis focusing on the content rather than the structure of the linguistic units. Accordingly, the last step of preparing our corpus was to generate a function-word-free version of the corpus. To attain this, we employed the NLTK library of Python once again and took the tokenized version of the corpus to provide a broader spectrum of the corpus versions. Therefore, we utilized Python lists to define the most common function words i.e. (articles, prepositions, and conjunctions) in the tokenized corpus. Then we create another list for all words not in the function words list. The latter list was joined again in one string using the *join()* function. The generated version was stored in a fourth folder named ASAWEC_content.

Quality Control

With the availability of advanced computer solutions that facilitate corpus compiling, new concerns emerged regarding the quality of the texts (Bodell et al., 2022) which may be sacrificed for the quantity. Regular users now can crawl hundreds of (born-digital) files that need not be retyped or processed to compile corpora. However, these files may be loaded with noise data and coding, and further annotation may cause more errors on them. Accordingly, it is envisaged that quality might be the coming concern in the corpus linguistics field. The emphasis on enhancing the quality of corpus files typically revolves around improving the annotation (Darģis et al., 2020), ensuring cleanliness, and eliminating duplicate content. However, most of these practices targeted corpora that are manually annotated e.g., learner corpora with error annotations. For files that are assembled from online sources, further post-compiling quality control measures are required to maintain corpus quality. These measures are controlled by research variables and data type. For the current corpus, all the procedures mentioned in this methodology section were implemented with accompanying quality and rigor in mind, yet more measures are taken as discussed below.

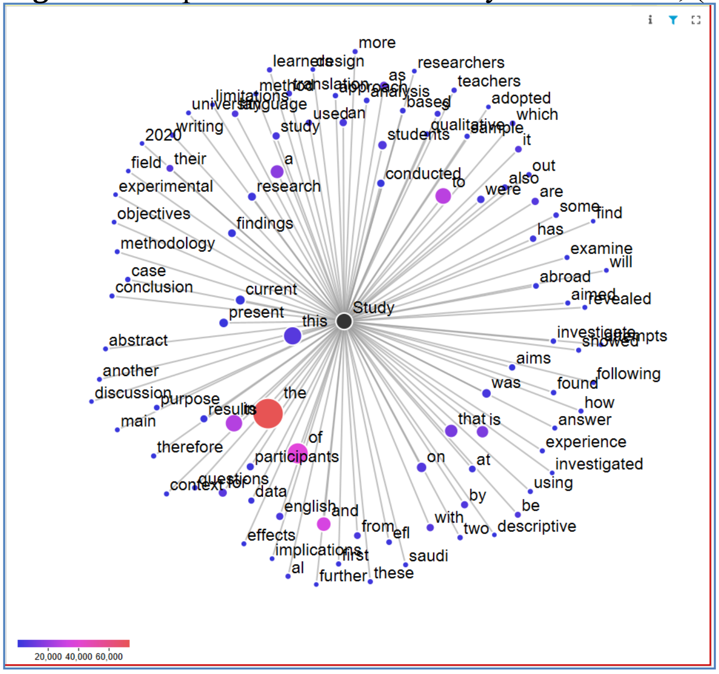
Content and structure review

The researchers utilized a peer review process to ensure the quality of the content in the corpus. Three university professors who specialize in Applied Linguistics, TESOL, and English Literature were selected to review the files. The files were distributed to them, and they were asked to visually scan the files for any extraneous data, such as Arabic characters, non-English symbols, extra spaces, or any other text that was not in English. Very few errors were detected, and these were promptly corrected. In addition, the corpus was cross-checked for indexing and structure. This involved reviewing the numbering system and ensuring that the file numbers corresponded accurately with the original PDF files. The sequence of the numbering system was also checked to guarantee the accuracy of the recorded metadata.

Corpus test

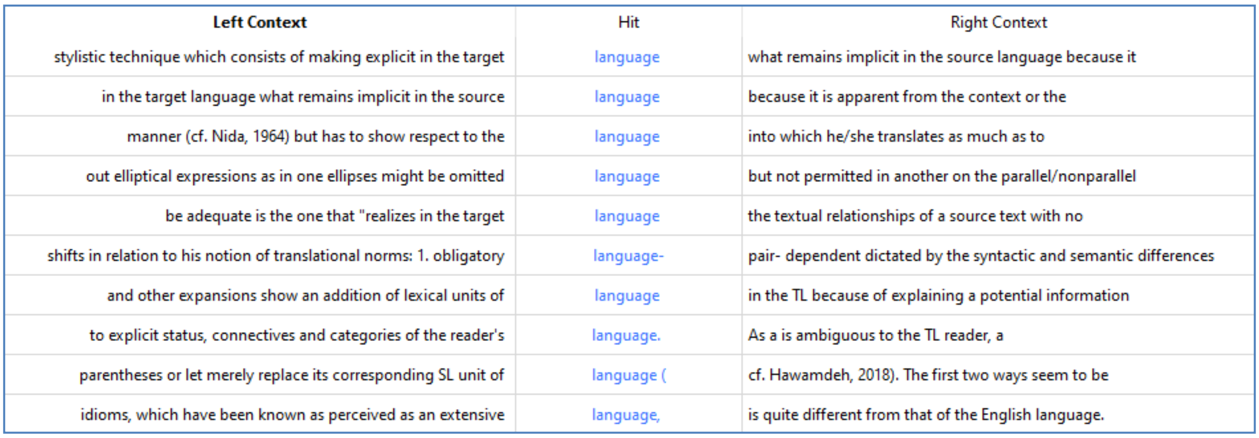
The corpus was then uploaded to two software packages which have been used widely in recent years. They are #LancsBox X (Brezina and Platt, 2023), and AntConc V 4.2.4 (Anthony, 2023). The corpus was tested for common analysis methods performed in corpus linguistics such as Wordlists, KWIC, collocations, Ngrams, keywords, etc. The test results of the GraphColl feature of the content word version of ASAWEC are displayed in Figure 2.

Figure 2. Graphcoll of the word Study in ASAWEC, (content word version)



GraphColl is a feature in #LancsBox software that shows the collocate of a word in visual representation in terms of direction and strength of the collocation. Figure 3 presents another example of the coprus test by AntConc for the KWIC of the word *language*.

Figure 3. KWIC for the word language by AntConc



Left Context	Hit	Right Context
stylistic technique which consists of making explicit in the target	language	what remains implicit in the source language because it
in the target language what remains implicit in the source	language	because it is apparent from the context or the
manner (cf. Nida, 1964) but has to show respect to the	language	into which he/she translates as much as to
out elliptical expressions as in one ellipses might be omitted	language	but not permitted in another on the parallel/nonparallel
be adequate is the one that "realizes in the target	language	the textual relationships of a source text with no
shifts in relation to his notion of translational norms: 1. obligatory	language-	pair- dependent dictated by the syntactic and semantic differences
and other expansions show an addition of lexical units of	language	in the TL because of explaining a potential information
to explicit status, connectives and categories of the reader's	language.	As a is ambiguous to the TL reader, a
parentheses or let merely replace its corresponding SL unit of	language (	cf. Hawamdeh, 2018). The first two ways seem to be
idioms, which have been known as perceived as an extensive	language,	is quite different from that of the English language.

Further analyses were conducted on the corpus to generate initial results and statistical information for the purpose of the corpus documentation.

Documentation

Documentation is an essential procedure in corpus compiling as it accounts for providing sufficient information for the prospective users about the corpus. It is also important because it retains the information that were deleted from the files for cleaning purposes. Accordingly, a documentation file incorporating most of the information provided in this article was prepared. This file includes corpus description, data collection and compilation plus instructions regarding copyrights and use.

Further, to provide detailed information about the corpus, especially that which was removed from the files during the compilation processes, we prepared a comprehensive metadata file in an Excel workbook. The workbook includes several sheets for the general and specific statistics of the corpus. The details of the whole corpus, each journal, and each file were provided including bibliographical information about the journals and classified information about each article's title, author/s, and publication date. It also includes statistical information about the articles such as the number of tokens, lemmas, types and the lexical density of each article.

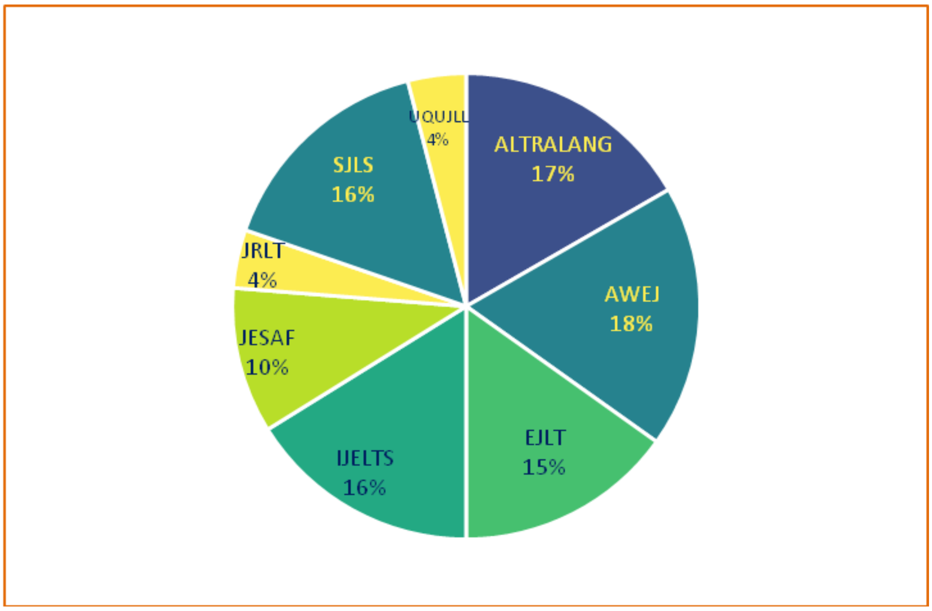
Statistics and Primary Results

The aim of this research project was to establish a specialized corpus for a published research article (Sanosi and Mohammed, 2024) and then to provide the resultant corpus for public use for non-commercial research purposes and make it an open source for further development and update. Through the compiling processes, the following results have been yielded.

Data source

As stated above, the corpus contains articles that we retrieved from the online archives of eight journals, six of which are open access while permissions from the publishers of the remaining two were acquired. Figure 4 displays the distribution of the articles and the journals.

Figure 4. Files per journal



We observed that there is a variance in the number of files extracted from each journal, this case was due to practical factors such as frequency and continuity of publication. Other quality control factors also had a role in this distribution as articles were removed for different factors such as the supremacy of Arabic (or French) content or the identity of the author. The statistics of the corpus regarding the files and journals are displayed in Table 2.

Table 2. Detailed file-per-journal statistics

Journal	Full Title	Files	Tokens	%	Indexing
ALTRALANG	Algerian Translation & Language Journal	33	144759	16.67%	1 - 33
AWEJ	Arab World English Journal	36	204549	18.18%	34 - 69
EJLT	The Egyptian Journal of Linguistic and Translation	30	198192	15.15%	70 - 99
IJELTS	International Journal of English Language and Translation Studies	32	157020	16.16%	100 - 131
JESAF	Journal of English Studies in Arabia Felix	20	85411	10.10%	132 - 151
JRLT	Journal of Research in Language and Translation	8	49982	4.04%	152 - 159
SJLS	The Saudi Journal of Language Studies	31	195732	15.66%	160 - 190
UQUJLL	Journal of Umm Al-Qura University for Language Sciences and Literature	8	54078	4.04%	191 - 198
Total		**198**	**1089723**	**100%**	**1 - 198**

Corpus Statistics

Roughly, ASAWEC is a proximately a one-million word-corpus. This size is relatively small compared to the contemporary corpora, nevertheless, it is deemed appropriate for regular research projects conducted by a single or two authors in a limited time. This is very common among the targeted audience of the project. Also, the corpus is in its initial stage, and we aspire to develop it soon to encompass other disciplines and maybe genres to include fields like ESP, EAP, and conference papers and research notes. The basic statistics of ASAWEC at this stage are provided in Table 3 below.

Table 3. Corpus Statistics

Item	Stat.
Files	198
Tokens	1089723
Types	41085
Type/token ratio (TTR)	0.04
Lemmas	36061
Lexical Density	0.839
Maximum file length	12578
Minimum file length	1073
Average file length	5504

The statistics reveal a relatively low lexical diversity with a TTR of 0.04. This indicates that there are relatively few unique words (types) compared to the total number of words (tokens) in the corpus. More precisely, it means that for every 100 words in the corpus, there are only 4 unique words. This suggests that the vocabulary used by the Arab scholars of English who wrote the corpus is limited and repetitive.

The fact that the genre is scientific articles, and the register is academic writing may partly explain the low TTR, as these types of texts tend to use specialized terminology and jargon, which can reduce lexical diversity. However, the TTR of 0.04 is still quite low, even for academic writing.

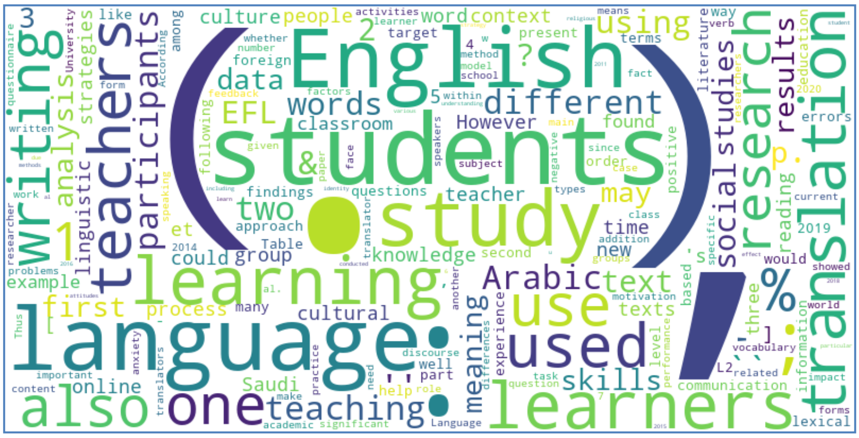
On the other hand, the high lexical density of the corpus (0.839) indicates that a large proportion of the words in the corpus are content words (nouns, verbs, adjectives, and adverbs), rather than function words (articles, prepositions, conjunctions, etc.). This is typical of academic writing, which tends to be more information-dense and less concerned with conveying social meaning than other types of writing.

Overall, the combination of a low TTR and a high lexical density suggests that the Arab scholars who wrote the corpus may have a limited vocabulary in English but are able to use the vocabulary they do know in a precise and focused way to convey scientific information. This initial result could be taken as a starting point for further research that will utilize this corpus for exploring lexical aspects of Arab scholars' writing.

Word frequency

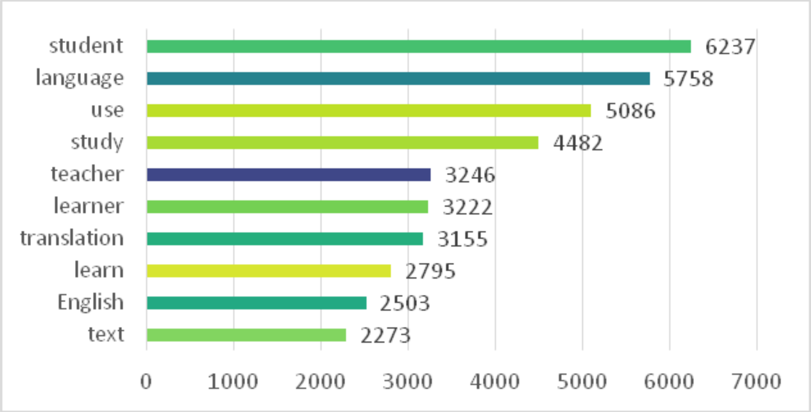
The low lexical diversity and high lexical density indicate that most of the words that are used in the corpus are used frequently. This fact can reflect the limited set of vocabulary due to the nature of this specialized corpus. As the disciplines and subdisciplines of this corpus are related to the English language and applied linguistics, it was anticipated that most of the words used would be related to English, ELT, linguistics, and literature. The word cloud in Figure 5 reveals that this anticipation was achieved.

Figure 5. Word cloud of the frequently used words in the corpus



As it is noted the most recurrent words in the corpus, the ones with larger font size, are related to applied linguistics. Frequent use of words such as *students, learning, learners, writing, and teaching* confirm this hypothesis. Additionally, the corpus incorporates words likely related to discourse analysis such as *discourse, text, meaning,* and *context* while there is a frequent use of research-related vocabulary such as *study, results, participants*, and *analysis.* To give a specific picture of the most used words, we searched for the top ten used words in the corpus with their frequency of occurrence the search yielded the results shown in Figure 6.

Figure 6. Top ten frequent keywords in the ASAWEC



These results can provide an initial insight into the content of the corpus and thus help researchers to decide on adopting the corpus if it fits their research scope and question(s).

Availability and Access

The ASAWEC compilers commit to making the corpus available for research community. As some of the content of the corpus is restricted and was obtained under specific conditions, we make this corpus available under an Attribution-NonCommercial-NoDerivatives 4.0 International license. According to this license, researchers are free to use, share, and redistribute the corpus. However, they must give appropriate credit to the compilers, use the corpus only for non-commercial purposes, and make no derivatives to the corpus unless for personal use. There are no additional restrictions to the use of ASAWEC.

The corpus is published on *Figshare.com*[1] which is "a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner". It enables users to upload files of various formats and set access controls and licenses. It also provides DOIs that facilitate citation and tracking of research output.

Although the main files of the corpus are available in a friendly format on *Figshare.com*, researchers are advised to contact the compilers for any further inquiries regarding the corpus. This procedure is also taken to ensure that the data is being used appropriately and in accordance with the set license deeds.

Future perspectives and update plans

One feature to enhance the accuracy and rigorousness of corpora is the continuous update to remains relevant and up-to-date. Hunston (2002) considered continuous update a condition for the representativeness of a corpus stating that "any corpus that is not updated rapidly becomes unrepresentative" (P. 30). This is mainly because aspects of language, such as academic writing, are dynamic and new trends continue to emerge. Further, corpus compilers aspire to boost their corpus in terms of both quantity and quality. Considering this, several update plans for ASAWEC were proposed, the first of which is to be done in six months while the other plans will be conducted on a yearly basis. The improvement plans are established to consider the following updates.

Content enhancement

It is proposed to both expand and structure and broaden the scope of the corpus by adding more research articles on various disciplines within humanities as academic writing is not exclusive to English applied linguistics and literature. This plan is

[1] ASAWEC is available at https://doi.org/10.6084/m9.figshare.24187461

proposed to enlarge the corpus and attract more researchers from different fields. This plan is the first to be executed and completed in a period of not more than one year.

Annotation and tagging

Advanced annotation and tagging are also proposed to make the corpus more versatile for linguistic research purposes. Suggested annotations include advanced annotation schemes such as discourse markers, metadiscourse resources, and semantic roles. Nowadays, several corpus linguistics and NLP applications incorporate advanced annotation tools that can annotate corpora making them more appropriate for various linguistic analyses. While these tools are accessible to researchers, preparing different versions of ASAWEC with different annotations will make it more convenient.

Access

The corpus will remain open-source and free for researchers under the stated license. Moreover, it is planned that after updating the corpus content, it will then be hosted on a special website with a user-friendly interface. It is also proposed that research and analysis tools to be integrated on the website to facilitate querying the corpus for specific linguistic features and topics. The researchers are open to collaboration with institutions and individuals who have similar projects and are open to collaborate to achieve these aims which need collective resources and cooperation.

Suggestion for use

ASAWEC was compiled bearing in mind that it should be available for a wide range of queries in academic writing by Arab scholars. Accordingly, several versions were generated to be suitable for different research. These versions are deemed proper for different research queries in discourse analysis, however, a few suggestions for the use of ASAWEC can be presented here.

The content word version is suitable for searching for vocabulary use, word frequency, and keywords. There is also a lemmatized version to search for stance, attitude, and other searches that focus on semantic aspects more than on syntactic ones. The full version can serve for searching for linguistic features in general. Examples for use include searching for collocation, lexical bundles, discourse and metadiscourse markers, and conjunctions.

Conclusion

The aim of this paper was to report and document the processes and procedures followed in compiling the nucleus of a proposed corpus for academic writing by Arab scholars. The aim of the project was to bridge the gap in this strand as there are very few corpora for academic written English in the Arab world that do not proportionate with the rich production of Arab scholars in this field. We discussed and detailed the processes we followed for data collection, processing, and corpus compiling. The main statistics and initial results of the procedures were presented and discussed. We further explained the metadata reporting and presentation and explained the availability and access options. Finally, we acknowledged the potential limitations of the project at this stage and explained our future plans for updating and enhancing the corpus.

This project is envisaged to enrich the field of corpus linguistics through conducting in-depth linguistic analysis and research. The corpus is deemed to serve as a valuable resource for studying the nature of Arabs' academic discourse and language use patterns. However, this aim is unachievable without further participation from researchers in the field by using the corpus, giving constructive feedback on its structure and content, and even actually participating in developing its subsequent versions. Participation in this endeavor would be invaluable in advancing academic research in our field.

References

Akeel, E. S. (2014) A corpus-based study of modal verbs in academic writing of English native speakers and Saudis: Theses in pursue of academic degree of Master's in Applied Linguistics, Reading, 72 p.

Allan, R., Shaw, I. and Shaw, M. (2023). Building a corpus of written tasks of Swedish national tests in English: Motivation, method, and research applications, *Nordic Journal of English Studies*, 22 (2), 128154. https://doi.org/10.35360/njes.821

Almohizea, M. (2017). The compilation process of (COLTLC): A learner corpus, *International Journal of Language and Linguistics*, 4 (4), 223–231.

Alotaibi, H. (2017). Arabic-English parallel corpus: A new resource for translation training and language teaching, *Arab World English Journal*, 8 (3), 319–337. https://dx.doi.org/10.24093/awej/vol8no3.21

Anthony, L. (2022). *AntFileConverter* (Version 2.0.2) [Computer Software], Waseda University, Japan, available at: https://www.laurenceanthony.net/software/antfileconverter/ (Accessed 07 October 2023)

Anthony, L. (2023). *AntConc* (Version 4.2.4) [Computer Software], Waseda University, Japan, available at: https://www.laurenceanthony.net/software (Accessed 07 October 2023)

Atkins, S., Clear, J. and Ostler, N. (1992). Corpus design criteria, *Literary and Linguistic Computing*, 7(1), 1–16. https://doi.org/10.1093/llc/7.1.1

Baker, P. (2010). *Sociolinguistics and corpus linguistics*, Edinburgh University Press, Edinburgh, Scotland.

Baker, P., Hardie, A. and McEnery, T. (2006). *A glossary of corpus linguistics*, Edinburgh University Press, Edinburgh, Scotland.

Bird, S., Klein, E. and Loper, E. (2009). *Natural language processing with python*, O'Reilly Media, Inc.Sebastopol, CA, USA.

Blecha, J. (2012) Building specialized corpora: Thesis in pursue of the academic degree of Master's in English Language and Literature, Masaryk, 159 p.

Bodell, M., Magnusson, M. and Mutzel, S. (2022). From documents to data: A framework for corpus quality, *Scoius: Sociological Research for Dynamic World*, 8, 1–15. https://doi.org/10.1177/23780231221135523

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*, Cambridge University Press, Cambridge, UK.

Brezina, V. and Platt, W. (2023). *#LancsBox X* [Computer Software]. Lancaster University, available at: https://lancsbox.lancs.ac.uk/ (Accessed 07 October 2023)

Collins, L. (2019). *Corpus linguistics for online communication*, Routledge, London, UK.

Cox, C. and Newman, J. (2020). Corpus annotation. In Paquot, M. and Gries, S. (eds.), *A practical handbook of corpus linguistics*, Springer, Cham, Switzerland, 25–49.

Crawford, W. and Csomay, E. (2016). *Doing corpus linguistics*, Routledge, London, UK.

Darģis, R., Auziņa, I., Levāne-Petrova, K. and Kaija, I. (2020). Quality-focused approach to a learner corpus development, *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020*), Marseille, France, 392–396. DOI: 10.13140/RG.2.2.13826.43207

Dunn, J. (2022). *Natural language processing for corpus linguistics*, Cambridge University Press, Cambridge, UK,

Fuentes, A. (2009). A case study corpus for academic English written by NNS authors, in Gómez, P. and Pérez, A. (eds.), *A survey of corpus-based research*, Spanish Association of Corpus Linguistic (AELINCO), Madrid, Spain, 1101–1114.

Gilquin, G. (2020). Learner corpora, in Paquot, M. and Gries, S. (eds.), *A practical handbook of corpus linguistics*, Springer, Cham, Switzerland, 283–304.

Guerra, J. and Smirnova, E. (2023). How complex is professional academic writing? A corpus-based analysis of research articles in 'hard' and 'soft' disciplines, *Vigo International Journal of Applied Linguistics* (20), 149–184. DOI: 10.35869/vial.v0i20.4357

Hunston, S. (2002). *Corpora in applied linguistics*, Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9781139524773

Jamalzadeh, M. and Tabrizi, H. (2020). Academic vocabulary in tourism research articles: A corpus-based study, *Journal of Language and Discourse Practice*, 1(2), 23–42. DOI: 10.14744/ldpj.2020

Kubler, S. and Zinsmeister, H. (2015). *Corpus linguistics and linguistically annotated corpora*, Bloomsbury Publishing, London, UK.

Lemmenmeier-Batinić, D. *Spoken language corpora: Approaches for facilitating linguistic research*, Dissertation in pursue of of Doctor of Linguistics. Zurich. 2023. 39 p.

Liu, D. (2022). Using corpora for learning academic writing: A systematic review, *The thirty-first International Symposium on English Language Teaching*, English Teachers'

Association-Republic of China (ETA-ROC), Taipei, Taiwan.

McEnery, T. and Wilson, A. (2001). *Corpus linguistics: An introduction*, Edinburgh University Press, Edinburgh, Scotland.

Meyer, C. (2023). *English corpus linguistics: An introduction*, Cambridge University Press, Cambridge, UK.

Sanosi, A. B. and Mohammed, A. (2024). A corpus-based analysis of Arab scholars' use of interactional metadiscourse markers. *International Journal of English Language and Literature Studies*, 13(2), 188-200. https://doi.org/10.55493/5019.v13i2.5006

Sanosi, A. B. (2022). The use and development of lexical bundles in Arab EFL writing: A corpus-driven study, *Journal of Language and Education*, 8 (2), 108–123. https://doi.org/10.17323/jle.2022.10826

Sanosi, A. B. and Mohammed, A. (2024). A corpus-based analysis of Arab scholars' use of interactional metadiscourse markers. *International Journal of English Language and Literature Studies*, 13 (2), 188–200. https://doi.org/10.55493/5019.v13i2.5006

Sinclaire, J. (1991). *Corpus, concordance, collocation*, Oxford University Press, Oxford, UK.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*, Language Science Press, Berlin, Germany.

Toriida, M.-C. (2016). Steps for creating a specialized corpus and developing an annotated frequency-based vocabulary list, *TESL Canada Journal*, 34 (11), 87–105. http://dx.doi.org/1018806/tesl.v34i1.1255

Utkina, T. (2021). Teaching academic writing in English to students of economics through conceptual metaphors. *The Journal of Teaching English for Specific and Academic Purposes*, 9 (4), 587–599. https://doi.org/10.22190/JTESAP2104587U

All authors have read and approved the final manuscript.

Все авторы прочитали и одобрили окончательный вариант рукописи.

Conflicts of Interest: the authors have no conflict of interest to declare.

Конфликты интересов: у авторов нет конфликта интересов для декларации.

Abdulaziz B Sanosi, PhD in Applied Linguistics, Lecturer, College of Science and Humanities, Prince Sattam bin Abdulaziz University, Hawtat Bani Tamim, Saudi Arabia.

Abuelgasim Sabah Elsaid Mohammed, PhD in Applied Linguistics, Assistant Professor, College of Science and Humanities, Prince Sattam bin Abdulaziz University, Hawtat Bani Tamim, Saudi Arabia.