International Journal of Educational Sciences and Arts (IJESA)
Vol. (4), No. (4)
IJESA
April 2025
المجلة الدولية للعلوم التربوية والآداب
الإصدار (4)، العدد (4)

The Saudi Learners of English Corpus (SLEC): Design, Research Potential and Applications

Ahmed Alhassani
Assistant Professor, PhD in Applied linguistics, King Abdullah Air Defense College, Saudi Arabia
alhassani2002@gmail.com

Abstract

The design and use of learner corpora is a rapidly developing branch of corpus linguistics. Learner corpora have the signal merit of allowing for the use of evidence-based methods in applied linguistics. This article introduces the Saudi Learners of English Corpus (SLEC), composed of writing by undergraduate students in Saudi universities. The SLEC is presented as the first Saudi written English as a Foreign Language (EFL) corpus that will be made eventually publicly available. It comprises 175,592 words, collected from EFL learners in Saudi Arabia, all of whom have studied English for nine years in Saudi public schools. The corpus includes data produced by 741 students. Their proficiency level ranges between beginner and intermediate. The corpus is designed to include a variety of metadata which describes features of the texts and the learners. The article presents the contents and the design criteria of SLEC, discussing in detail the rationale for the corpus, the participants involved, the corpus size, the materials included, the method of data collection, corpus metadata and architecture. Pedagogical implications and potential future research are also addressed.

Keywords: English as a foreign language, Learner corpus, Learner corpus design, Saudi learners of English.

1- Introduction

Learner corpora are increasingly being used in linguistic research in areas such as Language Teaching and Learning and Applied Linguistics, as well as for other purposes such as Error Analysis, Language Materials Designing, and the building of learner dictionaries. The number of learner corpora has increased rapidly in the last two decades, as have studies based on these corpora, such as those by Pravec (2002), Granger (2004), Nesselhauf (2004), Wen (2006), Granger et al. (2013), Díaz-Negrillo and Thompson (2013), and Granger and Dumont (2014), which suggests how important corpora have become in language learning research and how valuable a data resource they provide. Learner corpora are defined as "a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration" (McEnery and Wilson 2001, p. 32). In simpler terms, Gilquin & Granger (2015) describe them as "electronic collections of writing or speech produced by foreign or second language learners", assembled in a principled way and guided by a specific research objective. These collections of texts, which can be written or spoken, may contain mistakes and incorrect use of the language in question, compared to an idealized native speaker's output.

Learner corpora are regarded as a relatively new addition to the wide range of existing corpus types (Nesselhauf, 2004), with Granger (2008, p. 259) suggesting that learner corpus research "began to emerge as a discipline in its own right in the late 1980's/early 1990's". Learner corpora have been used by researchers for the purpose of Contrastive Interlanguage Analysis (Granger 1998) to investigate a wide range of instances of overuse, underuse, and misuse of many different aspects of learner language at different levels, including lexis, discourse, and syntax (Granger, 2003). In addition, they help researchers, teachers, material designers and dictionary designers to identify the interlanguage errors caused by first language (L1) transfer, learning strategies, and overgeneralization of L1 rules. These errors can happen at

the level of words, phrases, or language structures (Granger, 2003b; Nesselhauf, 2004). Moreover, learner corpora can be a useful resource to measure learners' improvement in various aspects of the target language (Buttery and Caines, 2012; Nesselhauf, 2004), which in turn can be used for pedagogical purposes in creating teaching materials that are more appropriate to learners' proficiency levels.

However, a lack of developing freely available learner corpora for Saudi Learners of English should be addressed to fill the gap in the research on English language learning and teaching in Saudi Arabia. The corpus presented in this article, the Saudi Learners of English Corpus (SLEC), thus responds to two previously unmet needs, namely it represents a publicly available resource[1], that contains samples of written, rather than spoken Arabic, and represents the English specifically of Saudi learners of English. Thus, this article aims to report the compilation of SLEC. The participants were 741 undergraduate students in Saudi universities who studying English as a foriegn language (EFL). Upon the completion of the corpus, corpus analytical methods were employed to conduct preliminary research, such as investigating basic corpus information and word frequencies.

2- Literature Review

There are many learner corpora which focus on English as the target language, such as the Advanced Learner English Corpus (ALEC), and only a few others which focus on languages other than English (Gilquin & Granger, 2015). This focus on English stems from the fact that it is the most spoken language worldwide (Seals and Shah, 2017) and it is also considered as the language of science (Ammon, 2007). However, most of the English produced worldwide is in fact generated by non-native speakers (Crystal, 2003). As a result, linguists and SLA researchers should focus more on the language of non-native speakers and learners in their research. Such focus would

[1] The corpus will be available publically eventually, but for the time being it can be available upon request from the researcher.

allow a shift in EFL and ESL curriculum design from an approach based on intuition to one that depends on authentic language use. Nonetheless, issues related to accuracy and adequacy in use of English by learners and non-native speakers are reported to create challenges for learner corpus designers.

In terms of Saudi context, over the previous decade or so, researchers have been working on compiling their own learner corpora of English. However, to the best of my knowledge and having reviewed the existing literature on learner corpora in Saudi EFL contexts, it appears that the only Saudi written EFL corpora in existence are very small corpora compiled mainly by individual researchers and PhD students to fulfill their private research needs. The only Saudi EFL corpus that could be found in the literature and which is available to researchers is the Saudi Learner Corpus (SLC), which is an official part of the Louvain International Database of Spoken English Interlanguage (LINDSEI), an ongoing project that includes spoken English produced by learners from eleven L1 backgrounds. The SLC is a corpus of informal interviews with higher intermediate-to-advanced EFL learners (Algouzi, 2014). Therefore, it is clear that the availability of written Saudi EFL corpora so far lags behind that of other L1 backgrounds. Consequently, the rationale behind the compilation of SLEC can be explained in terms of the lack of EFL Saudi learner corpora and the need to create a purpose-built corpus of English written by Saudi EFL learners. However, since the Saudi learners' first language is Arabic, it could be argued that there are other learner corpora where Arabic is the underlying L1, and where some of the participants are Saudi learners, so there is no need to compile a specific corpus for Saudi learners. This argument might apply, for example to the BUiD Arab Learner Corpus (BALC), which is a collaboration between the British University in Dubai and the University of Birmingham, the Qatar Learner Corpus (QLC), which is a spoken corpus, and the Longitudinal Database of Learner English (LONGDALE), which covers various L1 backgrounds including Arabic. This claim can be refuted by arguing that Arabic is a language that is used in different spoken

vernaculars, which are often considered as constituting different dialects, across the Arab world. Therefore, although the common background is Arabic, the differences in the way people across these countries use the language is clearly noticeable. In addition to that, the educational system and the way they treat English is different across these countries. English is spoken as a foreign language in some Arab countries (e.g. in Saudi Arabia and Egypt) and as a second language in others (e.g. in the UAE). And interlanguage factors mean that learners tend to make certain errors based on their L1 background, more specifically their informal L1background (Swan and Smith, 2001). Thus, research based on a cohort with a specific variety of the L1 (e.g. Saudi Arabic) should be more useful for pedagogical purposes.

In the rest of this article, I present the SLEC in detail, starting with the principles and criteria that were considered in the design of this corpus and moving onto the process of collecting the learner corpus texts and the metadata and then a detailed description of the corpus. Finally, the article is concluded by stating the pedagogical implications of this corpus and future work.

3- The Saudi Learner English Corpus

In this section, the principles and criteria that were considered in the design of this corpus is reported in some details, including collecting the metadata of the texts and the learners, the methods of recruiting participants, and the task used to collect the data.

3.1 Designing the learner corpus:

One of the key issues in corpus studies is the creation of the corpus itself. A random collection of heterogeneous learner data does not qualify as a learner corpus (Granger, 2002, p. 9). Rather, learner corpora should be compiled according to strict design criteria, some of which are the same as for native corpora (Granger, 2002). According to Conrad (2002, p. 77), determining the criteria that will guide corpus design, such as size of the corpus, types of texts included, number of texts, the

sampling procedure, etc., is crucial in order to achieve reliable results. In the following sections, the most important theoretical and practical considerations in corpus building and criteria for designing a corpus will be discussed in more detail.

3.2 Purpose of the corpus:

Before researchers start to design and compile corpus materials, they must set up clearly and exactly the purpose of the corpus, that is, what questions the corpus is supposed to provide answers for. The design criteria and the corpus construction will be guided by this purpose. Based on previous studies in learner corpora, two main categories can be generally identified, the first one is corpora which are meant to be used by a wide audience of users for broad aspects of research (public purpose), and the second kind are corpora which are meant to be used by a specific group of users to study specific aspects of language (specific purpose). Both kinds of corpora have their own special characteristics in their design and content. A corpus can be designed to serve one or more purposes such as language learning/teaching, material development, error analysis, descriptive analysis, translation and so on. For instance, the purpose of the the International Corpus of Learner English ICLE (Granger, 1998 and Granger et al, 2002) is "to make use of advances in applied linguistics and computer technology to effect a thorough investigation of the interlanguage of the foreign language learner" (Granger, 1993, p.57). Specific purposes corpora could examine the role of age and gender in learner language use, or to explore learners' lexico-grammatical and phraseological competence, or to record lexical uses, as is the case with the Bilingual Speech Corpus for French and German Language Learners (Fauth et al., 2014). SLEC, for its part, was compiled with objective of adding to the current body of research on Saudi learners of English, and to allow for future investigations of Saudi learner English, which could be of great help for dictionary compilers and material designers, as well as teachers and learners.

3.3 Size:

Size is a controversial issue in corpus development as it plays a significant role regarding the notion of representativeness as the result of which a corpus-based study could be generalized to the population of language learners.

The size of a given corpus is usually decided by the purpose for which it is intended. Sinclair (1991) suggested that "A large corpus of many millions of words is possible and should keep on growing" (p.18). To get useful empirical evidence regarding word use and collocation behavior, building a large corpus of many millions of words is useful. However, it is believed that a smaller homogenous corpus that features a high-quality design is more valuable than a larger corpus (Granger, 1993). Koester (2010) believes that smaller, but more specialized corpora allow researchers to take a deep look into patterns of language in a particular setting as there is a strong connection between the corpus and the contexts in which the texts in the corpus were produced, as compared to large corpora which are made of a mix of different texts types.

When it comes to learner corpora, Granger (2003b) argues that:

"Size is obviously a relative notion. A corpus of 200,000 words is big in the SLA field where researchers usually rely on much smaller samples but minute in the corpus linguistics field at large where recourse to mega corpora of several hundred million words has become the norm rather than the exception" (p. 465).

However, she believes that large learner corpora would be "a major asset in terms of representativeness of the data and generalizability of the results" (Granger, 2004:125).

But generally, does the size of a corpus really matter? According to Sinclair (2005), absolute size is not the most important consideration, and the size of the corpus should rely on two factors: "the kind of query that is anticipated from users and the

methodology they use to study the data" (p.10). Thus, there is no specific size the corpus should be as each corpus is built to address particular needs, but at the same time a learner corpus should be representative of the language of the learner being studied. However, the difficulty of collecting data from specific learners may force researchers to minimize the size of the corpus especially in the first stage of collecting data.

With respect to SLEC the intended size was to be over 200,000 words to be representative, however due to the fact that 45% of the particpants were beginners and the numbers of words each particpaint produce did not reach the expected number of words assigned for each one decressed the size of corpus.

3.4 Representativeness:

Corpus representativeness is arguably one of the most significant issues in corpus design and has been discussed in many studies (Sinclair 1991; Biber et al. 2002; Sinclair 2005). Sinclair (2005) argues that "Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen." (p.2). According to Leech (1991: 27), a corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety. Biber (2002) also points out that: "The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research." (p.246). SLEC was designed to be representative of the written English of Saudi undergraduate EFL learners.

Primarily, representativeness depends upon how balanced the corpus is, i.e. the range of text categories included in the corpus. Balance is an important issue in corpus creation (Hunston, 2002; Nelson, 2002) and refers to "the weighting between the different sections in a corpus" (Kennedy, 1998, p. 62). Similar to representativeness, the acceptable balance of a corpus is decided by its intended use. A balanced corpus

usually includes a wide range of text genres which are supposed to be representative of the language under consideration. According to Sinclair (2005) corpus designers "should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components." Thus, corpus designers should work to achieve a well-balanced corpus. In short, McEnery and Hardie (2011) believe that the measures of representativeness and balance are matters of degree.

With regard to SLEC, it could be argued that the data are taken from only one region in Saudi Arabia, thus it cannot be considered representative. However, this can be refuted by two points: first is that all learners in Saudi Arabia speak Arabic as their first language and secondly, they go through the same education system which applied everywhere is Saudi Arabia. In addition, the data where taken from two different cities in the region.

3.5 Metadata:

Metadata is defined as "data about data" (Burnard, 2005: p.30). Thus, it is a collection of information that depicts or "documents" (Granger, 2002) the corpus data. When designing a corpus, it is crucial to have this metadata as an important part of the corpus. Without metadata, it would be difficult to identify different patterns of linguistic behavior in naturally occurring samples of language as metadata "restores and specifies… context, thus enabling us to relate the specimen to its original habitat" (Burnard, 2005: p.31).

The question then arises of which variables should be included in the metadata of a learner corpus. Metadata, according to Granger (2008), can be classified along two major dimensions relating to characteristics of the learners who have produced the data (such as "age", "gender", "nationality"), and characteristics of the tasks they were requested to perform ("text mode", "text genre", "country of production", etc.).

These features can be used as determinants to search any subset of the corpus data or to conduct comparisons between different groups of learners or texts.

Based on the above discussion, the following guidelines summarize the criteria and metadata used to design SLEC:

Table 1: Corpus-building criteria

Criterion	Corpus characteristics
Purpose	The corpus allows for the investigation of different aspects of the written English produced by a specific group, namely undergraduate Saudi learners of English as a Foreign Language.
Language	English
Size	200,000 tokens, with a plan to increase the number
Target language	It contains data from a single language, namely English produced by Saudi EFL Learners.
Text dates	2019
Text location	Saudi Arabia, western region
Text type	Academic writing
Learners' proficiency level	Beginner and Intermediate
Material mode	Written
Task type	Essays
Material genre	argumentative, narrative, and descriptive

Table 2: Metadata elements used in in designing SLEC

Learner variables		Text variables	
1	Age	1	Text genre
2	Gender	2	Text mode
3	Nationality	3	Text medium
4	Native language	4	Text length
5	Number of years learning English	5	Year of production
6	Number of years spent in English-speaking countries	6	Country of production
7	Level of study	7	Timing
		8	Reference use
		9	Dictionary use

Thus, the corpus is designed to include a variety of metadata which describes features of the texts and the learners. The job of these features is to work as determinants to

explore any subset of the corpus data or to make comparisons between different groups of texts or learners. The corpus is thus marked up with 16 metadata variables. Seven are related to the learners and nine are related to the texts.

So, as shown in Tables 1 and 2, the purpose of designing this corpus is to investigate the use of different aspects of the English language used by a specific group of learners, namely undergraduate Saudi learners of English as a foreign language. The intended target size of this corpus is 200,000 words in the first stage of development. The proficiency level of the learners ranges between beginner and intermediate as they are undergraduate students whose level of English is expected not to be that high. As there was limited time for data collection, advanced learners were not included due to the fact that it would be difficult to find many advanced learners of English in Saudi Arabia and it would require more time and more university visits. In terms of L1 background, all learners share the same L1, which is Arabic. In addition, the corpus contains written data only, and includes three different material genres: argumentative, narrative, and descriptive.

3.6 Participants:

The learners who participated are 741 learners of English as a Foreign Language (EFL) rather than as a Second Language (ESL). Regarding their L1 background, learners are from only one mother tongue background, Arabic. The learners' proficiency predominantly falls in the beginner-intermediate range. They were 1st year Saudi undergraduate students of English as a foreign language who have studied English for nine years in Saudi public schools. The age range of these participants was between 18 and 24 years old. In terms of gender, the data were produced by (439) males' participants and (302) females' learners. Based on their performance on the Oxford Quick Placement Test, the results of only two groups will be compared, namely group A2 (beginner) with group B1 (lower intermediate), see appendix A.

The participants were asked to provide some demographic information such as native language, gender, years of studying English, number of years they have attended English classes, and other languages they speak fluently. In addition, the participants also had to indicate their age group: 1) Age group I (18-22), 2) Age group II (22-26), and 3) Age group III (26-30). These learner profiles thus provide researchers with information which allows comparisons across different sections of the corpus.

3.7 Task:

In terms of task type, essays were used to collect the data as it is the most preferable task type in the literature (Granger and Gilquin, 2015; Kennedy, 1998). Three genres were included: argumentative, narrative and descriptive. All texts were written in class as part of writing activities with no help from further resources or reference works such as grammar books, monolingual dictionaries, or bilingual dictionaries. Participants were asked to produce between 150 and 300 words as an average length for each text in a timed setting (lasting approximately 40 minutes) on different topics (See appendix B). All the data were elicited in one country, Saudi Arabia, more specifically in the western region of Saudi Arabia (Taif and Almadinah) as knowing the place of the production of data could help researchers to identify any differences in the language use of learners from different cities. All the texts were collected over a three-month period between April and June 2019.

4- Data Gathering

The corpus data was not taken from previously existing materials; instead, a particular methodology was designed to carefully collect and manage the corpus data. This methodology includes designing tasks, setting the standards for converting the hand-written texts into electronic form, measuring the consistency between transcribers of written data, and storing and managing the data through creating a database and generating different types of files automatically. The methodology including all these processes is explained in the following sections.

4.1 Collecting the Data:

The corpus contains one type of medium: materials written by hand. Guidelines were created to illustrate the steps the researcher (or his representative) would follow in collecting the data. Data collection took place during one main session that was repeated with each group of students, typically representing one class, at each educational institution. During each session, which was expected to last for about 1 hour, an assignment was distributed, and procedures were explained to the participants. The assignment consists of five parts as follows:

1. Information sheet, which includes a brief outline of the project, its benefits, data collection procedures, and participation in the research.
2. Consent form, in which the participant agrees that the purpose of the research has been explained to them, that they have understood the purpose of the study, and that they understand that their participation in this study is voluntary and that they are free to withdraw from the study at any time, without giving a reason and without consequence; they also agree that their answers, which they have given voluntarily, can be used anonymously for research purposes.
3. Learner and task metadata (information about the participant and the task being performed).
4. The Oxford Quick Placement Test, which places learners on the Common European Framework of Reference for Languages (CEFR) to better understand how the use of some linguistic feature can vary according to participants' proficiency according to the CEFR, see appendix C. The participants were given 20 minutes to finish the test. Following the test guidelines, scores were then converted into approximately equivalent bands to those of the CEFR, as can be seen in Table 3. Consequently, a distinctive feature of SLEC is that it can later be divided into sub-corpora according to these proficiency bands.

Table 3: Breakdown of Participants' L2 profiecncy

CEFR Band	Grouping
A2	Lower-level
B1	Intermediate
B2	Intermediate
C1	Upper intermediate
C2	Advanced

5. The Task, which includes writing one text (narrative or argumentative or descriptive) in class.

Each of the learners signed a consent form which stated that the data collected could be published and used in relevant future research. As already indicated, participants were both male and female. However, the education system in Saudi Arabia is organized based on single gender classes, thus genders do not mix. For this reason, it would have been so difficult if not impossible for a male researcher to go into a female university during their working hours to collect data. Therefore, it was necessary to recruit female representatives to help collect the required data. Two female representatives were asked to sign a consent form confirming that all data they collected would be saved in a secure place until they were handed on to the researcher at the end of collection process. The form stated that the representatives were not allowed to keep any part of the data in any medium and would not share any information they might know about the learners or their materials with any third party. The researcher also got permission from the institution from which the corpus data was collected for him or his representative to meet students and collect the corpus materials. Most of the participants were all motivated to contribute to this project due to its importance to research into the teaching of English in Saudi Arabia. As a result, they were not paid for their participation as they were happy to participate voluntarily.

After the researcher or his representative introduced the research, learners could ask any question about the research, its purposes, or their participation before signing the

form. Then the task was distributed with an explanation of how to complete it. First, participants were given The Oxford Quick Placement Test to be completed in 20 minutes. Then they were given the writing task, which was timed (40 minutes), and the learners were not allowed to consult any language references (e.g., dictionaries, grammar books) while writing their essays. Table 4 shows an example of the instructions for one of the tasks

Table 4: Task Instructions

The text: Do you agree or disagree with the following statement: Watching TV is bad for children. Write at least 250 words.
Time: 40 minutes.
Place: in class.
Language references: during this task you are NOT allowed to use any reference tools such as dictionaries or grammar books.
Medium of writing: writing these texts is by hand on the sheets provided by the researcher; two pages are provided for each text.

4.2 Collecting the Corpus Metadata:

The learner profile questionnaire of the International Corpus of Learner English (Granger, 1993, 2003b; Granger et al., 2010) was used to collect the metadata. However, some modifications were made in order to suit current purposes. Some questions which are related to learners' relatives were omitted, such as father's and mother's mother tongue, as were questions related to primary and secondary education, medium of instruction, etc. As indicated above, 17 elements were collected as the corpus metadata, eight related to the learner and nine were associated with the text.

4.3 Computerizing the Corpus:

Corpora containing hand-written texts require further work to convert them into an electronic form to make them readable by most language processing tools. Transcribing such hand-written with no standards, specifically when it is done by more than one transcriber, can result in differences in the final production, as many

items may be omitted or added during the transcription process and thus change the results of the corpus analysis. For the conversion process, the researcher used specific standards as described below.

4.4 Transcribing the Data:

As the corpus data is all derived from hand-written texts, specific standards were created in order to achieve a high level of consistency in transcription. Those standards address issues such as how to handle struck out words, or a doubtful form of a character, or unknown words or phrases. Two transcribers, the researcher and one volunteer English teacher who works as teacher of English at Taif University, performed the transcription based on a number of agreed-upon standards (see Appendix D). The agreed-upon standards were also discussed and revised by transcribers prior to the task to ensure consistency in transcribing, and additional reviews were performed throughout the transcription process when transcribers came across uncertain points.

4.5 Annotation:

Another decision to be made in the process of compiling the learner corpus is to annotate and error-tag the raw texts. So, all the transcribed essays were uploaded to Sketch Engine (Kilgarriff et al., 2014) to compile it into a corpus. Sketch Engine was preferred for use in this study because uploading the researcher's own corpora is possible, and it has many advantages and functions such as the availability of the Corpus Query Language (CQL), which can be used to input complex search queries for specific structures or collocations, and automatic Part-of-speech (POS) tagging. The corpora were then automatically tagged using the tagset named "TreeTagger Part-of-speech (POS) tagset with Sketch Engine Modifications".

5- Corpus Description

As described earlier, the SLEC is a corpus comprising a collection of written materials from learners of English in Saudi Arabia. The corpus includes 212,033 tokens (running words and punctuation) and 175,592 running words, 1,156 documents, produced by 741 students. Table 5 presents the basic frequency counts and statistics of the corpus.

Table 5: frequency counts and statistics of the corpus

COUNTS

Tokens	212,033
Words	175,592
Sentences	11,693
Documents	1,156

The corpus content based on the metadata of the learners shows that those learners are all Saudi undergraduate learners of English for whom Arabic is their L1. The age of the learners ranges between 18 and 22. When it comes to gender, 51.6 % of participants are female while 48.4 % are male (Fig. 1). The data collected from males amounted to 102,699 words; and from females to 109,334 words. The number of years they spent on learning Arabic ranges from 6 to 9 years and the number of years they had spent in an English-speaking country is 0 years. As for proficiency levels, 594 documents were produced by lower intermediate and 562 by beginner Tokens 210,858 Words 175,588 Sentences 11,688 Documents 1,156 Figure (1) Distribution of tokens based on gender of the learners. 181 learners of English. The data collected from beginners were 93,340 tokens, while 117,518 tokens were collected from lower intermediates. Figure (2) shows the current contents of the corpus per level of proficiency.

International Journal of Educational Sciences and Arts (IJESA)
Vol. (4), No. (4)
IJESA
April 2025
المجلة الدولية للعلوم التربوية والآداب
الإصدار (4)، العدد (4)

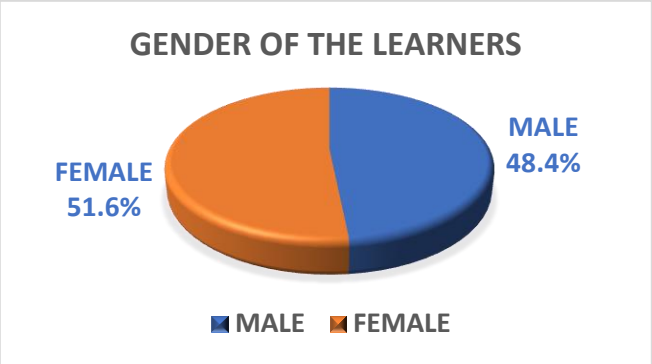

Fig. 1 Distribution based on gender of the learners

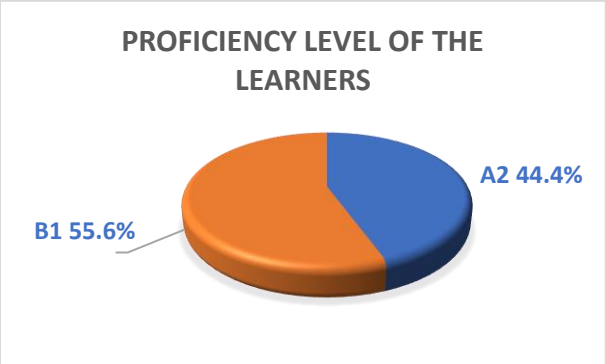

Fig. 2 Distribution based on proficiency level of the learners

Figure (3) shows the current contents of the corpus per level and per gender.

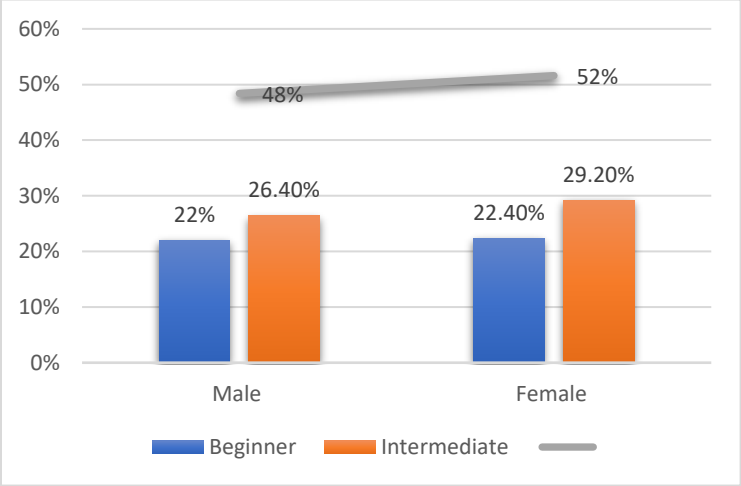

Fig. 3 Distribution per level and per gender of the learners

The texts produced by the learners were categorized into three different genres, Argumentative 33.6%, Descriptive 36.9% and Narrative 29.6%. All corpus data were produced in the classroom in the western region of Saudi Arabia in a timed setting with no reference works or dictionary use. The average length of the texts is 151 words.

International Journal of Educational Sciences and Arts (IJESA)
Vol. (4), No. (4)
IJESA
April 2025
المجلة الدولية للعلوم التربوية والآداب
الإصدار (4)، العدد (4)

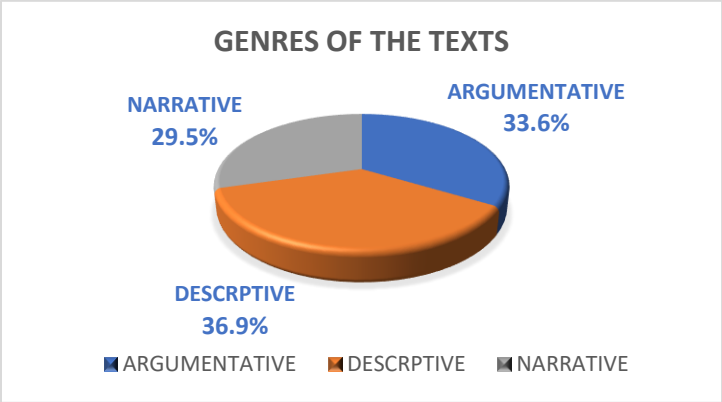

Fig. 4 Distribution based on genres of the texts

Using the word list function in Sketch Engine, the most frequent words were retrieved. The cut-off point was 300 occurrences, that is the only words that occurred at least 300 times were included. 87 words were extracted that occurred more than 300 times. Table 6 below shows the most frequently used words in the corpus and it is obvious that function words such as I, the, and, to and my take the lead. The next step will be to divide the corpus into sub-corpora and analyze what words are overused or underused.

Table. 6 The most frequent words in SLEC

	Word	Freq		Word	Freq		Word	Freq
1	I	7099	31	he	746	60	should	454
2	the	6971	32	on	730	61	first	454
4	and	6535	33	this	724	62	has	439
5	to	5552	34	when	720	63	if	419
6	my	3729	35	like	713	64	them	415
7	in	3521	36	will	708	65	their	407
8	a	3329	37	very	705	66	get	404
9	it	2902	38	but	701	67	food	403
10	is	2833	39	about	697	68	am	396
11	of	2476	40	be	685	69	want	392
12	you	2099	41	then	676	70	internet	388
13	for	1762	42	your	674	71	make	386
14	we	1726	43	also	669	72	eat	385
15	that	1550	44	not	598	73	bad	375

16	was	1451	45	went	596	74	up	369
17	with	1277	46	family	576	75	take	360
18	at	1198	47	some	569	76	as	357
19	have	1145	48	after	568	77	think	347
20	there	917	49	good	551	78	best	338
21	because	910	50	lot	543	79	place	326
22	go	909	51	from	540	80	watch	317
23	can	905	52	or	523	81	use	315
24	do	903	53	many	502	82	restaurant	314
25	so	865	54	day	494	83	see	314
26	are	806	55	children	492	84	saudi	310
27	time	776	56	our	484	85	things	309
28	they	772	57	life	468	86	university	305
29	me	771	58	one	468	87	by	300
30	people	749	59	all	460			

Concordance lines are usually presented in the KWIC format in which words, phrases, or combinations of words are clearly displayed in the middle of concordance lines (Sinclair, 2003). A close semantic analysis of these concordances lines can provide access to many important language patterns in texts and the types of words or phrases associated with the word in question and its pattern of meaning. An example of concordance lines from SLEC resulting from searching for the word *time* is given in Figure (5). In this way, concordance data presented in the KWIC format makes it easy for learners for instance to see what words occur immediately before and after the keyword.

International Journal of Educational Sciences and Arts (IJESA)
Vol. (4), No. (4)
IJESA
April 2025
المجلة الدولية للعلوم التربوية والآداب
الإصدار (4)، العدد (4)

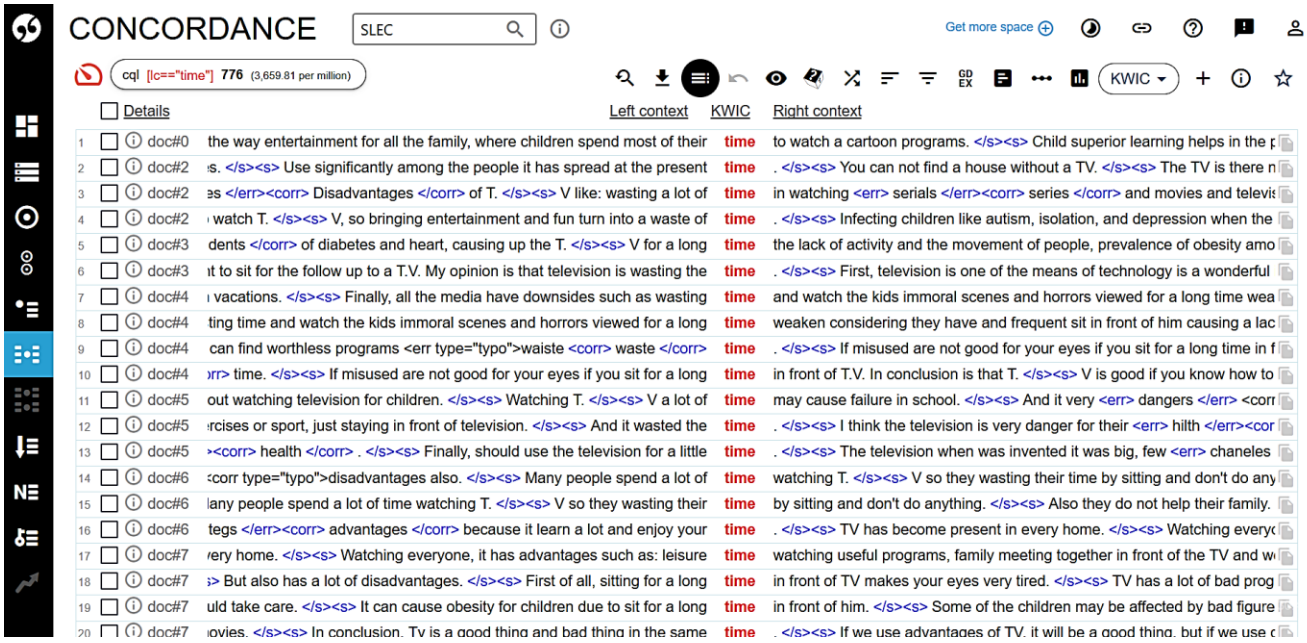

Fig. 5 Screenshot of Sketch Engine showing concordance lines of time in SLEC

Lexical bundles (Biber, Finegan, Johansson, Conrad, & Leech, 1999) is another common lexical analysis in Corpus Linguistics is of chunks, which are recurrent strings of words used together repeatedly. The N-gram tool in Sketch Engine offers the function of automatically produces frequency lists of sequences of tokens. The n-gram size was set between three- and four-words with a minimum of five occurrences.

Table. 7 The most frequent Lexical bundles in SLEC

	word	Freq		word	Freq
1	a lot of	413	51	around the world	43
2	with my family	216	52	I would like	43
3	I go to	209	53	talk about my	42
4	I went to	175	54	go to bed	42
5	we went to	153	55	and my family	42
6	in Saudi Arabia	138	56	When I was	41
7	I want to	117	57	went with my	40
8	I do not	117	58	this is my	40
9	go to the	114	59	in your restaurant	40
10	went to the	103	60	It was a	40

11	you have to	100	61	we went to the	40
12	In my opinion	99	62	I would like to	40
13	one of the	96	63	I wake up at	40
14	in the morning	93	64	lot of people	39
15	to go to	87	65	have a lot	39
16	I wake up	79	66	I went with	39
17	wake up at	71	67	to quit smoking	38
18	in my life	67	68	go to sleep	38
19	There are many	67	69	do my homework	38
20	advantages and disadvantages	65	70	and go to	38
21	is one of	63	71	My name is	38
22	my daily routine	57	72	I think it	38
23	in the world	57	73	I get up	38
24	has a lot	56	74	I went with my	38
25	the most important	54	75	the Internet is	37
26	my family and	54	76	lot of time	37
27	I have a	54	77	because it is	37
28	is one of the	54	78	and i I	37
29	there is a	52	79	I eat my	37
30	in front of	52	80	there are many	36
31	a long time	52	81	of Saudi Arabia	36
32	has a lot of	52	82	is the most	36
33	to talk about	51	83	is my daily	36
34	go to university	51	84	and this is	36
35	you want to	50	85	I take a	36
36	when I was	49	86	I like to	36
37	do not have	49	87	I hope you	36
38	i I was	48	88	one of the most	36
39	to be a	47	89	have a lot of	36
40	of the most	47	90	a lot of time	36
41	get up at	47	91	me and my	35
42	I will talk	47	92	I think the	35
43	I did n"t	47	93	is my daily routine	35
44	will talk about	46	94	you do n"t	34
45	Saudi Arabia is	46	95	with my friends	34
46	would like to	45	96	take a shower	34
47	is the best	45	97	it is very	34
48	We went to	44	98	it is a	34
49	After that I	44	99	for a long	34
50	I will talk about	44	100	lot of things	33

The tool returned the result for the three- to four-word clusters that appear five or more times. Table 7 lists the top 100 lexical bundles in the writing of Saudi learners of English. Most of them are grammatical groups. N-gram produces words being used together which are not necessarily meaningful multi-word units and they lack any syntactic unity or semantic integrity... O'Keeffe et al. (2007, p. 61) pointed out the bundles generated by corpus software might consist of highly-frequent fragmentary word groups, syntactically incomplete but meaningful strings and semantically and pragmatically fixed expressions. The lexical bundles in the writing of Saudi learnerds need to be further investigated. In addition, further research can be done in conjunction with native speaker corpus.

6- Conclusion and Future Work

This paper described the significance, corpus design and data compilation process for constructing the Saudi Learners of English Corpus. As there are no publicly available corpora of writings by Saudi learners of English, SLEC stands to make a valuable contribution to the field of Saudi EFL studies. Although it is time-consuming in terms of data entry, the effort required to create SLEC should be worthwhile in the long term and the corpus will benefit not just the current researcher, but researchers and English teachers and students in Saudi Arabia and beyond.

English as foreign language research may benefit from SLEC as it provides authentic data produced by learners at different proficiency levels as well as extensive metadata. In addition, these data can be used to carry out contrastive interlanguage analysis by comparing them with native speaker data to uncover patterns of overuse, underuse, and misuse in learner lexis and discourse according to different criteria such as gender, proficiency level, years of learning English and genres. Moreover, SLEC data could be useful for Saudi English language teachers anxious to improve their teaching performance. Teachers can use the SLEC data to set up different

exercises and activities and increase learners' awareness by providing access to frequent learning errors.

In the future, the plan is to maintain and enlarge the corpus by adding data from learners with more different proficiency levels or even to add a spoken part. Another potential future project would involve researching the process of second language acquisition using the corpus data to investigate different aspects such as collocational patterns, error patterns, and spelling errors specific to Saudi learners of English, to identify what is particularly difficult for them, and to put special emphasis on these points into the design of teaching materials.

Acknowledgement

I would like to express my very great appreciation to Prof. Dorothy Kenny and Iker Erdocia for their valuable and constructive suggestions during the planning and development of this research work.

References

Algouzi, S. (2014). Discourse markers in Saudi English and British English: A comparative investigation of the use of English discourse markers. Unpublished PhD Thesis. University of Salford.

Ammon, U. (2007). Global Scientific Communication: Open Questions and Policy Suggestions. In: Carli, A., Ammon, U. (Eds.), Linguistic inequality in scientific communication today. John Benjamins, Amsterdam/Philadelphia, (pp. 123–133).

Biber, D., Conrad, S., & Leech, G. N. (2002). Longman student grammar of spoken and written English. Harlow: Longman.

Burnard, L. (2005). Metadata for corpus work. In M. Wynne (Ed.), Developing Linguistic Corpora: A Guide to Good practice (pp. 30–46). Oxford, UK: Oxbow Books.

Buttery, P. and A. Caines (2012) Normalising Frequency Counts to Account for 'opportunity of use' in Learner Corpora, in Developmental and Crosslinguistic Perspectives in Learner Corpus

Research, Y. Tono, Y. Kawaguchi, and M. Minegishi, Editors. 2012, John Benjamins: Amsterdam. p. 187-204.

Conrad, S. (2002). Corpus Linguistic Approaches for Discourse Analysis. Annual Review of Applied Linguistics; Cambridge, 22, 75–95.

Crystal, D. (2003). English as a Global Language. Ernst KlettSprachen.

Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards the future. In Díaz-Negrillo, A., Ballier, N., & Thompson, P. (Eds). Automatic Treatment and Analysis of Learner Corpus Data (pp. 9–30). Amsterdam, the Netherlands: Benjamins.

Fauth, C., Bonneau, A., Zimmerer, F., Trouvain, J., Andreeva, B., Colotte, V., Fohr, D., Jouvet, D., Jügler, J., Laprie, Y., Mella, O., & Möbius, B. (2014). Designing a Bilingual Speech Corpus for French and German Language Learners: A Two-Step Process. In: Proceedings of the LREC 2014, International Conference on Language Resources and Evaluation (pp. 1477–1482). Reykjavik, Iceland: European Language Resources Association.

Gilquin, G., & Granger, S. (2015). Learner language. In: Douglas Biber & Randi Reppen, The Cambridge Handbook of English Corpus Linguistics, Cambridge University Press: Cambridge 2015, p.418-435.

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), English language corpora: Design, analysis and exploitation (pp. 57–69). Amsterdam, the Netherlands: Rodopi.

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

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising Synergy. CALICO Journal, 20(3), 465–480. JSTOR.

Granger, S. (2004). Practical Applications of Learner Corpora. In Lewandowska-Tomaszczyk, Barbara (ed) Practical Applications in Language and Computers, Frankfurt: Peter Lang, 291–301.

Granger, S. (2004b). Computer Learner Corpus Research: Current Status and Future Prospects. In U. Connor & T. Upton (Eds.) Applied Corpus Linguistics: A multidimensional perspective (pp. 123–145). Amsterdam, the Netherlands: Rodopi.

Granger, S. (2008). Learner Corpora. In A. Ludeling & M. Kyto (Eds.), Corpus linguistics: An international handbook (pp. 259–275). Berlin, Germany: Walter de Gruyter.

Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), Computer learner corpora, second language acquisition and foreign language teaching (pp. 3–33). Amsterdam, the Netherlands: Benjamins.

Granger, S. (2003). The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. TESOL Quarterly, 37(3), 538–546.

Granger, S. (2012). How to use foreign and second language learner corpora? In Mackey, A. &Gass, S.G. (Eds.), A Guide to Research Methods in Second Language Acquisition (pp.7-29). Basil: Blackwell.

Granger, S., Gilquin, G., & Meunier, F. (2013). Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011). Presses universitaires de Louvain.

Granger, S. & Paquot, M. (2010). The Louvain English for Academic Purposes dictionary. In S. Granger & M. Paquot (Eds.) eLexicography in the 21st century: New applications, new challenges. Proceedings of eLEX2009. Cahiers du Cental 7. Louvain-la-Neuve: Presses universitaires de Louvain, 87-96.

Granger, S., & Dumont, A. (2014). Learner corpora around the world. UCL Centre for English Corpus Linguistics. Accessed November, 30, 2019.

Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge: Cambridge

University Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., & Suchomel, V. (2014). The sketch engine: Ten years on. Lexicography 1: 7–36.

Koester, A. (2010). Building small specialised corpora. In A. O'Keeffe and M. McCarthy (Eds.) Routledge Handbook of Corpus Linguistics (pp.66-79). London: Routledge.

Leech, G. (1991). The state of the art in corpus linguistics. In Aijmer, K. and Altenberg, B. (eds.), English Corpus Linguistics: Studies in honour of Jan Svartvik. Longman, London, pp. 8 – 29.

McEnery, T., & Hardie, A. (2011). Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). Corpus Linguistics: An Introduction. Edinburgh: Edinburgh University Press.

Nelson, M. B. (2000). Corpus-based study of the lexis of business English and business English teaching materials. University of Manchester. Retrieved from https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.488069

Nesselhauf, N. (2004). Learner corpora: Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), How to use corpora in language teaching (pp. 125–152). Amsterdam, the Netherlands: Benjamins.

Pravec, N. A. (2002). Survey of learner corpora. ICAME Journal, 26(1), 8–14.

Seals, C. A., & Shah, S. (2017). Heritage Language Policies around the World. London: Routledge.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.

Swan, M., & Smith, B. (2001). Learner English: A teacher's guide to interference and other problems. Cambridge University Press.

Thompson, P. A. (2005). Spoken language corpora. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice (pp. 59–70). Oxford, UK: Oxbow Books.

Wen, Q. (2006). Chinese learner corpora and second language research. Paper presented at the 2006 International Symposium of Computer-Assisted Language Learning, Beijing, China.

International Journal of Educational Sciences and Arts (IJESA)
Vol. (4), No. (4)
IJESA
April 2025
المجلة الدولية للعلوم التربوية والآداب
الإصدار (4)، العدد (4)

Appendices

Appendix A

Measuring the proficiency level of the learners.

NO.	Number	Gender	The score	Proficiency level
1	4003045	Male	32	B1
2	3612998	Male	35	B1
3	4028803	Male	37	B1
4	4001949	Male	33	B1
5	4002164	Male	37	B1
6	4001866	Male	44	B2
7	4001593	Male	32	B1
8	4029489	Male	36	B1
9	4028944	Male	36	B1
10	4002676	Male	38	B1
11	4000736	Male	32	B1
12	4000352	Male	33	B1
13	4001308	Male	32	B1
14	4005037	Male	39	B1
15	4003286	Male	42	B2
16	4002186	Male	30	B1
17	4000025	Male	35	B1
18	4003195	Male	33	B1
19	3929365	Male	33	B1
20	4002095	Male	32	B1
21	3611727	Male	34	B1
22	4000847	Male	37	B1
23	4028181	Male	39	B1
24	4001878	Male	32	B1
25	4028797	Male	34	B1
26	4001078	Male	32	B1
27	4001449	Male	33	B1
28	4000091	Male	32	B1
29	4002629	Male	39	B1
30	4000410	Male	38	B1
31	4000625	Male	39	B1
32	4000021	Male	36	B1
33	4003425	Male	34	B1
34	4002792	Male	33	B1
35	4001596	Male	33	B1
36	4001790	Male	38	B1
37	3900316	Male	39	B1

38	3705117	Male	33	B1
39	4001151	Male	32	B1
40	4001691	Male	36	B1
41	4002300	Male	36	B1
42	4028858	Male	36	B1
43	4003292	Male	33	B1
44	4028503	Male	32	B1
45	4000305	Male	32	B1
46	4000726	Male	35	B1
47	4000207	Male	33	B1
48	4029316	Male	33	B1
49	4002845	Male	34	B1
50	4002098	Male	33	B1
51	4003230	Male	32	B1
52	4002171	Male	39	B1
53	4003332	Male	33	B1
54	4002616	Male	34	B1
55	4000586	Male	37	B1
56	4000262	Male	35	B1
57	4003305	Male	33	B1
58	4000113	Male	35	B1
59	4001488	Male	33	B1
60	3901390	Male	34	B1
61	4003521	Male	34	B1
62	4000619	Male	18	A2
63	4000833	Male	20	A2
64	3414070	Male	22	A2
65	4001932	Male	20	A2
66	3902092	Male	26	A2
67	4005953	Male	19	A2
68	3903173	Male	15	A1
69	4002597	Male	18	A2
70	4003296	Male	18	A2
71	4028322	Male	24	A2
72	4000724	Male	22	A2
73	3902824	Male	21	A2
74	4002184	Male	19	A2
75	4000645	Male	25	A2
76	4002987	Male	21	A2
77	4000472	Male	20	A2
78	4002824	Male	23	A2
79	4000720	Male	18	A2
80	3900088	Male	25	A2
81	4002603	Male	10	A1

82	4003409	Male	18	A2
83	4000700	Male	23	A2
84	4001307	Male	19	A2
85	4006171	Male	20	A2
86	3902479	Male	20	A2
87	3929612	Male	21	A2
88	4001299	Male	17	A1
89	4028650	Male	22	A2
90	4028809	Male	21	A2
91	3900107	Male	21	A2
92	4028456	Male	26	A2
93	4028527	Male	15	A1
94	4028480	Male	22	A2
95	4000919	Male	24	A2
96	4029672	Male	19	A2
97	4005194	Male	21	A2
98	4000790	Male	23	A2
99	4004949	Male	19	A2
100	4028769	Male	26	A2
101	4001602	Male	19	A2
102	4028467	Male	17	A1
103	4028312	Male	20	A2
104	3904182	Male	20	A2
105	4029265	Male	21	A2
106	4000422	Male	26	A2
107	4000515	Male	18	A2
108	4000372	Male	20	A2
109	4028020	Male	19	A2
110	4006151	Male	18	A2
111	4028349	Male	26	A2
112	4002877	Male	15	A1
113	4002976	Male	20	A2
114	4015822	Male	19	A2
115	4012845	Male	20	A2
116	3002657	Male	25	A2
117	4008746	Male	24	A2
118	4026864	Male	22	A2
119	4003576	Male	23	A2
120	4008654	Male	22	A2
121	3611256	Male	20	A2
122	4028654	Male	26	A2
123	4004678	Male	20	A2
124	4001876	Male	22	A2
125	4007346	Female	23	A2

126	4001964	Female	22	A2
127	4028533	Female	18	A2
128	4029533	Female	20	A2
129	4001235	Female	22	A2
130	4007543	Female	20	A2
131	4001754	Female	26	A2
132	4025565	Female	19	A2
133	4009065	Female	20	A2
134	4001048	Female	15	A1
135	4001834	Female	22	A2
136	4019743	Female	21	A2
137	4002238	Female	19	A2
138	3929678	Female	25	A2
139	4001132	Female	21	A2
140	3611767	Female	20	A2
141	4000926	Female	23	A2
142	4028188	Female	18	A2
143	4018654	Female	25	A2
144	4082326	Female	20	A2
145	4001975	Female	21	A2
146	4021449	Female	22	A2
147	4002291	Female	25	A2
148	4012629	Female	24	A2
149	4018410	Female	20	A2
150	4000232	Female	21	A2
151	4000099	Female	17	A1
152	4009854	Female	23	A2
153	4002792	Female	20	A2
154	4028887	Female	18	A2
155	4011790	Female	20	A2
156	3902316	Female	22	A2
157	3708654	Female	20	A2
158	4012985	Female	26	A2
159	4022556	Female	19	A2
160	4032310	Female	26	A2
161	4027554	Female	22	A2
162	4013454	Female	22	A2
163	4028503	Female	19	A2
164	4000455	Female	19	A2
165	4000766	Female	22	A2
166	4000243	Female	15	A1
167	4025464	Female	18	A2
168	4012845	Female	21	A2
169	4032098	Female	38	B1

170	4013230	Female	39	B1
171	4022171	Female	44	B2
172	4003454	Female	38	B1
173	4032636	Female	38	B1
174	4043586	Female	39	B1
175	4003262	Female	36	B1
176	4034505	Female	37	B1
177	4034213	Female	37	B1
178	4024543	Female	37	B1
179	3906542	Female	39	B1
180	4016521	Female	43	B2
181	4024519	Female	37	B1
182	4024423	Female	38	B1
183	3426570	Female	38	B1
184	4015632	Female	36	B1
185	3905743	Female	37	B1
186	4015343	Female	37	B1
187	3903753	Female	36	B1
188	4036437	Female	33	B1
189	4009984	Female	36	B1
190	4021982	Female	36	B1
191	4000777	Female	38	B1
192	3902232	Female	35	B1
193	4009655	Female	34	B1
194	4010645	Female	34	B1
195	4032987	Female	38	B1
196	4036532	Female	33	B1
197	4025544	Female	37	B1
198	4015630	Female	38	B1
199	3904898	Female	36	B1
200	4032603	Female	42	B2
201	4018765	Female	33	B1
202	4023855	Female	32	B1
203	4018765	Female	38	B1
204	4019997	Female	33	B1
205	3902579	Female	36	B1
206	3929512	Female	34	B1
207	4001434	Female	36	B1
208	4028775	Female	38	B1
209	4022869	Female	35	B1

Appendix B

Titles of the essays used in the corpus

1- The most important technology and its advantages and disadvantages.

2- Inventions that changed our life.

3- How to keep healthy.

4- Advantages and disadvantages of traveling.

5- Internet and its advantages and disadvantages.

6- Advantages and disadvantages of watching TV.

7- Parents are the best teachers; do you agree or disagree?

8- Smoking should be banned in public; do you agree or disagree?

9- Advantages and disadvantages of using Mobile phones.

10- Advantages and disadvantages of Video games

11- Fast food should be banned in schools, do you agree or disagree?

12- What do you think is the right age of getting married and why?

13- Teaching sign language in public school, do you agree or disagree?

14- Advantages and disadvantages of using electrical cars.

15- Talk about one of the bad, or happy, or embarrassing or scary or dangerous experience you have had.

16- Is it ok to lie and what makes people lie?

17- Talk about your first day in the school, or university.

18- Talk about your first day in a foreign country.

19- Talk about your last holida.

20- Describe your country to someone visiting it.

21- Talk about your dreams and hopes.

22- Talk about a special gift you have received.

23- Write about your best friend.

24- Write about your house.

25- Talk about a person you know very well.

26- Write about the new Important changes in Saudi Arabia.

27- Talk about your life and daily routine.

28- How to quit smoking?

29- How to succeed in college?

30- Talk about your Favorite movie.

31- Talk about Your favorite place to relax in Saudi Arabia.

32- Talk about your mother or father.

33- Introduce yourself.

34- Talk about your favorite food and how to make it.

35- Describe your city.

36- Talk about a nice trip you did.

37- Talk about your favorite singer or actor.

38- Talk about your favorite hobby.

39- How does wedding look like in your country.

Appendix C

Oxford Quick Placement Tes

Oxford Quick Placement Test Section B

(Oxford University Press and University of Cambridge Local Examinations Syndicate)

There are 60 sentences. You have 20 minutes to finish this part. Good luck!

Question 1 – 5

• Where can you see these notices?

• For questions 1 to 5, mark one letter A, B or C on your Answer Sheet.

1. YOU CAN LOOK, BUT DON'T TOUCH THE PICTURES

A► in an office B► in a cinema C► in a museum

2. PLEASE GIVE THE RIGHT MONEY TO THE DRIVER

A► in a bank B► on a bus C► in a cinema

3. NO PARKING PLEASE

A► in a street B► on a book C► on a table

4. CROSS BRIDGE FOR TRAINS TO EDINBURGH

A► in a bank B► in a garage C► in a station

5. KEEP IN A COLD PLACE

A► on clothes B►on furniture C► on food

Question 6 –10

• In this section you must choose the word which best fits each space in the text below.

• For questions 6 to 10, mark one letter A, B, or C on your Answer Sheet

THE STARS

There are millions of stars in the sky. If you look (6).....the sky on a clear night, it is possible to see about 3000 stars. They look small, but they are really (7).....big hot balls of burning gas. Some of them are huge, but others are much smaller, like our planet Earth. The biggest stars are very bright, but they only live for a short time. Every day new stars (8).....born and old stars die. All the stars are very far away. The light from the nearest star takes more (9).....four years to reach Earth. Hundreds of years ago, people (10).....stars, like the North Star, to know which direction to travel in. Today you can still see that star. 6.

A► at B► up C► on

7.

A► very B► too C► much

8.

A► is B► be C► are

9.

A► that B► of C► than

10.

A► use B► used C► using

Question 11 - 15

• In this section you must choose the word which best fits each space in the texts.

• For questions 11 to 20, mark one letter A, B, C or D on your Answer Sheet.

Good smiles ahead for young teeth

Older Britons are the worst in Europe when it comes to keeping their teeth. But British youngsters (11).....more to smile about because (12).....teeth are among the best. Almost 80% of Britons over 65 have lost all or some (13).....their teeth according to a

International Journal of Educational Sciences and Arts, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJESA.2025.v4n4p5 E-ISSN 2976-7237

World Health Organisation survey. Eating too (14).....sugar is part of the problem. Among (15)....., 12-year-olds have on average only three missing, decayed or filled teeth. 11.

A► getting B► got C► have D► having

12.

A► their B► his C► them D► theirs

13.

A► from B► of C► among D►between

14.

A► much B► lot C► many D►deal

15.

A► person B► people C► children D►family

Question 16 - 20

Christopher Columbus and the New World

On August 3, 1492, Christopher Columbus set sail from Spain to find a new route to India, China and Japan. At this time most people thought you would fall off the edge of the world if you sailed too far. Yet sailors such as Columbus had seen how a ship appeared to get lower and lower on the horizon as it sailed away. For Columbus this (16).....that the world was round. He (17).....to his men about the distance travelled each day. He did not want them to think that he did not (18).....exactly where they were going. (19)....., on October 12, 1492, Columbus and his men landed on a small island he named San Salvador.

Columbus believed he was in Asia, (20).....he was actually in the Caribbean. 16.

A► made B► pointed C► was D► proved

International Journal of Educational Sciences and Arts, London Vol (4), No (4), 2025
https://doi.org/10.59992/IJESA.2025.v4n4p5 E-ISSN 2976-7237

17.

A► lied B► told C► cheated D► asked

18.

A► find B► know C► think D►expect

19.

A► Next B► Secondly C► Finally D►Once

20.

A► as B► but C► because D►if

Question 21 - 30

• In this section you must choose the word or phrase which best completes each sentence.

• For questions 21 to 40, mark one letter A, B, C or D on your Answer Sheet.

21. The children won´t go to sleep.....we leave a light on outside their bedroom.

A► except B►otherwie C► unless D► but

22. I´ll give you my spare keys in case you.....home before me.

A► would get B► got C► will get D► get

23. My holiday in Paris gave me a great.....to improve my French accent.

A► occasion B► chance C► hope D► possibility

24. The singer ended the concert.....her most popular song.

A► by B► with C► in D► as

25. Because it had not rained for several months, there was a.....of water.

A► shortage B► drop C► scare D► waste

26. I´ve always.....you as my best friend.

A► regarded B► thought C► meant D► supposed

27. She came to live here.....a month ago.

A► quite B► beyond C► already D► almost

28. Don´t make such a.....! The dentist is only going to look at your teeth.

A► fuss B► trouble C► worry D► reaction

29. He spent a long time looking for a tie which.....with his new shirt.

A► fixed B► made C► went D► wore

30. Fortunately,.....from a bump on the head, she suffered no serious injuries from her fall.

A► other B► except C► besides D► apart

Question 31 – 40

31. She had changed so much that.....anyone recognized her.

A► almost B► hardly C► not D► nearly

32. .....teaching English, she also writes children´s books.

A► Moreover B► As well as C► In addition D► Apart

33. It was clear that the young couple were.....of taking charge of the restaurant.

A► responsible B► reliable C► capable D►able

34. The book.....of ten chapters, each one covering a different topic.

A► comprises B► includes C► consists D►contains

35. Mary was disappointed with her new shirt as the colour.....very quickly.

A► bleached B► died C► vanished D►faded

36. National leaders from all over the world are expected to attend the.....meeting.

A► peak B► summit C► top D► apex

37. Jane remained calm when she won the lottery and.....about her business as if nothing had happened.

A► came B► brought C► went D►moved

38. I suggest we.....outside the stadium tomorrow at 8.30.

A► meeting B► meet C► met D►will meet

39. My remarks were.....as a joke, but she was offended by them.

A► pretended B► thought C► meant D►supposed

40. You ought to take up swimming for the.....of your health.

A► concern B► relief C► sake D►cause

Questions 41 – 45

• In this section you must choose the word which best fits each space in the texts.

• For questions 41 to 45, mark one letter A, B, C or D on your Answer Sheet.

CLOCKS

The clock was the first complex mechanical machinery to enter the home, (41)……..it was too expensive for the (42)…….person until the 19th century, when (43)…….production techniques lowered the price.

Watches were also developed, but they (44)…….luxury items until 1868, When the first cheap pocket watch was designed in Switzerland. Watches later became (45)……available, and Switzerland became the world´s leading watch manufacturing centre for the next 100 years.

41.

A► despite B► although C► otherwise D► average

42.

A► average B► medium C► general D► common

43.

A► vast B► large C► wide D► mass

44.

A► lasted B► endured C► kept D► remained

45.

A► mostly B► chiefly C► greatly D► widely

Questions 46 - 50

Dublin City Walks

What better way of getting to know a new city than by walking around it? Whether you choose the Medieval Walk, which will (46)…….you to the city as it was 1000 years ago, find out about the more (47)…….history of the city on the Eighteenth Century Walk, or meet the ghosts of Dublin´s many writers on The Literary Walk, we know you will enjoy the experience. Dublin City Walks (48)..……..twice daily. Meet your guide at 10.30 a.m. or 2.30 p.m. at the Tourist Information Office. No advance (49)……..is necessary. Special (50)……are available for families, children and parties of more than ten people.

46.

A► introduce B► present C► move D► show

47.

A► near B► late C► recent D► close

48.

A► take place B► occur C► work D► function

49.

A► paying B► reserving C► warning D► booking

50.

A► funds B► costs C► fees D► rates

Question 51– 60

• In this section you must choose the word or phrase which best completes each sentence.

• For questions 51 to 60, mark one letter A, B, C or D on your Answer Sheet.

51. If you´re not too tired we could have a……..of tennis after lunch.

A► match B► play C► game D► party

52. Don´t you get tired……watching TV every night?

A► with B► by C► of D► at

53. Go on, finish the dessert. It needs……up because it won´t stay fresh.

A► eat B► eating C► to eat D► eaten

54. We´re not used to…….invited to very formal occasions.

A► be B► have C► being D► having

55. I´d rather we…….meet this evening, because I´m very tired.

A► wouldn´t B► shouldn´t C► hadn´t D►didn´t

56. She obviously didn´t want to discuss the matter so I didn´t……..the point.

A► maintain B► chase C► follow D► pursue

57. Anyone……after the start of the play is not allowed in until the interval.

A► arrives B► has arrived C► arriving D► arrived

58. This new magazine is ……...with interesting stories and useful information.

A► full B► packed C► thick D► compiled

59. The restaurant was far too noisy to be……to relaxed conversation.

A► conducive B► suitable C► practical D► fruitful

60. In this branch of medicine, it is vital to ……..open to new ideas.

A► stand B► continue C► hold D► remain

Appendix D

Standards for the data transcription in the corpus.

1	The texts should be transcribed without any corrections and should remain authentic. No spelling mistakes or errors are corrected.
2	All the metadata variables should be excluded from the text body.
3	Any struck-out texts should be discarded.
4	If there is a correction above a non-struck out word, the correct form should be Transcribed.
5	The teacher's corrections and comments should be excluded
6	When there is a doubtful form of a character which cannot be transcribed, the closer possible form to the correct form is transcribed.
7	Inserting a new line (paragraph) only when it is clear.
8	Any identity information (e.g. learner's name, contacts, postal address, emails, etc.) should be replaced with # (personal information deleted).
9	Any text format, ornamentations, shape, illustration or underlined words or sentences drawn by the learner on the sheet should be excluded.