The Compilation Process of (COLTLC): A Learner Corpus

Monira Ibrahim Almohizea
King Saud University
Riyadh, Saudi Arabia

Abstract

*Corpus-based research has increasingly proved to be one of the rigorous evidence-based methods in the field of applied linguistics. Learner corpus-based applications in language learning and teaching–which is closely associated with research from within the Second Language Acquisition (SLA) tradition– has resulted in new strides in language learning and teaching. The present paper introduces Arabic-speaking **L**earner **C**orpus of students majoring in Arabic/English translation in the **C**ollege **O**f **L**anguages and **T**ranslation (COLTLC thereafter) at King Saud University in Riyadh. The available corpora of Arabic-speaking learners will be presented. Then, the gap in SLA pure approaches will be discussed to argue for the plausibility of learner-corpus-based studies and the need to compile COLTLC. The content and construction of COLTLC will also be presented. More specifically, the rationale, design, tagging, annotation and architecture of the corpus will be discussed in greater detail, as well basic descriptive for potential future research.*

Introduction

Corpus linguistics is a branch within applied linguistics that has become one of the dominant methods used to analyse language today. Learner corpus research is grounded in both, corpus linguistics and Second Language Acquisition (SLA) research (Granger, 2002).Corpus compilation as a process can be traced back to the 1960s, but after the seminal works of (Leech1992; Sinclair, 1991) this field started to expand. Research on corpora was mostly descriptive and based on written samples of English which mostly contributed to corpus-based approaches to grammar. The field of learner corpus research in particular, only started decades later. In the early 1990s corpora of on-native English speakers and learners of English were compiled as what we now know as the "learner corpus" (Granger 2002). McEnery and Wilson (2001) define a learner corpus as "a finite-sized body of machinereadable text, sampled in order to be maximally representative of the language variety under consideration" (p.32). Similarly and in more detail, Grangerdefineditas the electronic collections of authentic texts produced by English as a Foreign Language (EFL)and English as Second Language (ESL) learners of English in a principled way guided by a particular research objective. Such texts are made up of assembled texts (it could be spoken or written) which –as opposed to native-speaker's corpus, may contain inaccurate and erroneous use of English.

The importance of learner corpus research stems from the fact that English is the most spoken language worldwide (Seals and Shah, 2017) and it is also seen as the language of science (Ammon, 2007). More importantly, the majority of English produced worldwide is in fact generated by non-native speakers (Crystal, 2003). Therefore, the language of non-native speakers and learners should be the focal point for linguists, NLP researchers, SLA researchers and scientists alike. The reason is that it simply constitutes a crucial part of human communication. They also form a crucial element in EFL and ESL curriculum design and restructure– from being dependent on intuitions which must be treated with caution, to a one that is based on authentic language use. Nonetheless, this type of English production imposes several challenges. These are primarily related to accuracy and adequacy in use of English by learners and non-native speakers.

The purpose of this paper is threefold. First, it aims to survey existing Arabic-speaking learner corpora. Second, it sets out toargue that SLA research can greatly benefit from corpus-based methods. Third, the compilation of COLTLC motivated by an objective to explore use of prepositions by EFL learners will be presented in relation to how decisions were made, in addition to basic descriptives of the corpus.

2. SLA research meets corpus linguistics

SLA research has been traditionally heavily dependent on elicited experimental data, and it has been criticised for disfavoring naturally occurring data (Lozano and Mendikoetxea, 2013).

In terms of scope, research within SLA has been generally focused on the study of children language acquisition, longitudinal studies of one or two learners to gauge certain developmental patterns in SLA. According to Ellis (1994), the three types of traditional SLA research data are language use data, metalinguistic judgement tasks and self-report of introspective methods. Therefore, early works of employing learner corpus methods in SLA grew out of a concern of the heavy reliance on artificial over authentic language use and data. Along the same line, Mindt (1996) observed that EFL learners can hardly cope with English used by native speakers in real life. Learner corpora thus can provides a new hybrid type of data which can further understanding of some SLA research challenging notions such as developmental patterns of learning (see, e.g. Housen, 2002) and other mechanisms of language acquisition which traditional SLA approaches alone have explained rather insufficiently. The use of learner corpora can be seen− in a way to overcome some of the shortcomings of traditional SLA approaches.

One of the benefits of using natural occurring data is that learners' performance can be examined in large quantities. An aspect that elicited data would most probably lack. The merit of learner corpus research also lies in the fact that its tools and methods allow for the possibility of studying multiple linguistic features of inter/intra-learners variables from various L1 backgrounds comparatively or separately. Most importantly, corpora can be easily distributed and shared, so that the results of investigations can be verified and compared by different researchers. This in turn, allows researchers to generalise, cross-check and extrapolate their research findings more confidently. Learner corpus research tools and softwares with rich functions enable researchers to detect and explore a wide range of, stylistic, semantic phonological and syntactic phenomena, from various genres (spoken and written). Such research can make a huge contribution to research and development of applied linguistics as a whole.

Although corpora are considered one source of evidence, it basically complements, rather than replaces other introspection and elicitation data. One of the advantages of corpora is that they can provide quantitative data which intuitions (a deeply rooted notion in traditional SLA) cannot provide reliably. For some features in language frequency can be the most influential factor, and it is best captured in corpus-based methods. McEnery and Wilson (2001) argue that corpus is "the only reliable source of evidence for such features as frequency" (p. 12). However, they observe that the use of quantification in corpus linguistics goes beyond simple counting analysis. This essentially implies that corpus-based studies are both, qualitative and quantitative in nature.

According to Myles (2015), SLA researchers have been slow in taking advantage of learner corpora and their methodologies. However, a large body of research in SLA is now moving towards learner corpora for the types of tools, potentials and insights they can foster. Learner corpora are now used by researchers, teachers of English, and learners for various purposes. According to Leech (1998), corpus allows a systematic study of the learner's linguistic behaviour from the point of view of 'overuse' and 'underuse' compared to that of native speakers. Granger (1998) rightly argued that learner corpus research has widened the scope of SLA research as it treats the learner performance data altogether, rather than decontextualisng it as in traditional error analysis approaches.

This field has been successfully applied to a wide range of research arenas, particularly in SLA (e.g.Nesselhauf, 2005; Lozano, and Mendikoetxea, 2013; Zhao, 2014).It has also contributed and advanced methods in language teaching (Sinclair, 2004), Contrastive Interlanguage Analysis (CIA) (see e.g.Granger, 2002),and language testing (e.g.Alderson 1996).Learner corpora could be used to inform L2 lexicography, syllabus design, material design, data-driven learning and self-learning, and Farr (2008) goes further to argue that the use of learner corpora should be part of teacher education and training.These are just to name but a few of the possible research areas which can benefit from studies based on corpora. In terms of the objective for which COLTLC was compiled, a learner-corpus based study can be very promising to explore the use of prepositions by EFL learners.

3. Background on Arabic- speaking learner corpora

Over the last decade or so, researchers around the world have been compiling their own learner corpora of English. Considering the number of native Arabic speakers worldwide, relatively little work has been done on the compilation of Arabic L1-English L2 learner corpus. In spite of the fact that compiling a large-scale learner corpus cannot be easily achieved by the efforts of an individual researcher, learner corpus compilation in the Arab world has mainly been conducted by individual researchers and PhD students, except for few corpora which will be presented below. In this section, I will review the existing Arabic-speaking learner corpora. Of note, by this review, I do not intend to provide a comprehensive or anything like an exhaustive list of all the available Arabic-

speaking learner corpora. Researchers are constantly creating corpora to fulfill their research needs, and tracing all these corpora is almost impossible.

The aim of the review is to argue for the need to compile COLTLC to explore the use of prepositions by EFL leaners at KSU. This review is largely based on learner corpora listed in the Centre for English Corpus Linguistics (CECL) of the Université Catholique de Louvain (UCL) webpage: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html. In general, there are two major types of learner corpora. First: corpora that is purely designed for commercial purposes, i.e. for curriculum and textbooks design, as well as compiling dictionaries, but not for research purposes. Large-scale corpora are mainly commercial and sponsored by publishers. An example of which is The Cambridge Learner Corpus (CLC) of various L1 backgrounds of exam scripts. This is one of the largest corpora as it is comprised of 50 million words as of yet. It includes thousands of exam papers written by Cambridge ESOL test-takers from all over the world, including Arabic speakers. The other example is the Longman Learners' Corpus (LLC), which consists of 10 million words written by learners of English from about 160 different L1 backgrounds**.** Longman encourages students and teachers to send their essays and exam texts in exchange of the Longman Dictionary of Contemporary English to help them create LLC. Both, CLC and LLC are not available for the public since their use is restricted to the creation of pedagogical material for EFL learners by editorial staff.

The second type of corpora are mainly non-commercial[1]and are collected at universities or research centers by academics to further knowledge and understanding of various research objectives. In terms of L1 groups, some corpora included Arabic-speaking learners' corpus as a sub-corpus with many other sub-corpora of various L2 backgrounds, while others merely focused on Arabic L1 learners of English alone.

3.1. Exclusively Arabic L1 learner corpora

One of the available corpora that merely focused on learners from one particular country is the BUiD Arab Learner Corpus (BALC). This project is a collaboration between the British University in Dubai and the University of Birmingham. It consists of 1,865 texts written by university students, as well as secondary school students. It comprises 290,000 word tokens and 20,275 word types. The scripts were first hand written and then digitised to be incorporated into the corpus. According to Randall and Groom, (2009), this corpus was compiled to investigate spelling errors and the types of errors of Arabic-speaking learners in their free writing. In terms of availability, a copy of the current version of the corpus is available upon request.

Another collaboration of two universities learner corpora focusing on Arabic-speaking learners of English is Qatar Learner Corpus (QLC). This spoken corpus was created at the Modern Language Department of the Carnegie Mellon University. The material of the corpus was from interviews with 19 learners whose native language was Arabic, mostly speakers of the Qatari dialect[2]. A learner profile for each participant including L1, grade, nationality, gender, the rates of the participant's reading skills and language usage, as well as the average English level are included. Moreover, transcripts of all the interviews are available on the corpus website online and they contain linked utterances for each segment of the interview. Zhao and MacWhinney, (2010) used this corpus to study English tense-aspect acquisition.

3.2. Arabic L1 learner corpora as a sub-corpus

The other type of corpora included Arabic with many other L1 backgrounds (i.e., a sub-corpus). It is worth mentioning that the inclusion of Arabic greatly varies across these corpora. Some include substantial amounts of Arabic L1 data, while others only included one Arabic-speaking participant (e.g. the lang-8 corpus which will be discussed below).

The most prominent of all learner corpora is the International Corpus of Learner English (ICLE) which is a huge project compiled in collaboration with several universities lead by CECL at UCL in Belgium. It contains 3.7 million words of essays (mainly argumentative) written by advanced learners of English at the university level from learners representing 16 L1 backgrounds divided into sub-corpora. It can be regarded as the most principled accessible large-scale learner corpus of English. Unfortunately, Arabic-speaking learners are not yet included and according to the CECL website there is collaboration with a university in Morocco, but the work is yet to be completed.

[1] Some compilers might charge maintenance fees for the use of their corpora.

[2] Out of the 19 participants,two were not Qataris (one was Kuwaiti and the other was Jordanian)

The latest version of ICLE featuring a built-in concordancer was published in 2009 and is available upon request. ICLE also provides a reference corpus. The Louvain Corpus of Native English Essays (LOCNESS), which is comprised of exam essays written by native speakers of English (British and American students).

Another corpus that was also developed in CECL at UCL is the LONG itudinal Data base of Learner English (LONGDALE) which covers various L1 backgrounds including Arabic. This is a distinctive type of corpus in the sense that it covers both modes of learners' production (written and spoken) from a wide range of text types.
The participants in this corpus are 117 EFL undergraduate students in their first year at UCL, and the data was collected from them at three intervals over three years' period which makes it a valuable resource to study developmental patterns of learners.

As little attention has been paid to the spoken language produced by learners, CECL compiled the Louvain International Database of Spoken English Interlanguage (LINDSEI). This ongoing project includes spoken English produced by learners from eleven L1 backgrounds. It started after the creation of ICLE with a specific purpose to provide a spoken counterpart to it. It is comprised of oral data of picture description and interviews of upper-intermediate to advanced learners. So far the word count of the corpus reached 800,000 words. Similar to LOCNESS, a comparable reference spoken corpus of interviews with native speakers of English (LOCNEC) has also been compiled in parallel. The Arabic-speaking learner collaboration is in progress and the corpus is available on a CD.

Similar to the large-scale CLC in utilising exam scripts is the Educational Testing Service (ETS). This corpus is made up of 12,100 essays written by EFL learners as part of the standardized Test of English as a Foreign Language (TOEFL). The corpus was compiled in 2006 and 2007 by the TOEFLtest-takers.It included Arabicand 10 other L1 backgrounds. As stated in the website[3], more essays were added in 2014. Also included are the (topics) for the essays and information about the test takers' proficiency level. The corpus was developed with the specific objective of native language identification in mind. Therefore, it was divided into training data and testing data. Compilers of this corpus argue that it can be further exploited to support tasks and studies in the educational domain, including grammatical error detection and correction, and automatic essay scoring, in addition to a broad range of research studies (Blanchard, Tetreault, Higgins, Cahill, and Chodorow, 2013). Samples of this corpus are available online.

The Lancaster Corpus of Academic Written English (LANCAWE) is comprised of academic writing essays by non-native speakers of English enrolled in the pre-sessional programme and undergraduate courses at Lancaster University. Each participant contributes more than one piece of writing at different intervals (i.e. before, during and after taking the course) of different types of essays (descriptive, argumentative, etc.), which is again suitable for longitudinal developmental type of studies. According to Lee (2010),this corpus differs from ICLE in the sense that its participants are in an English-immersion environment. As is the case with ICLE, a small native speaker sub-corpus was compiled in parallel to it as a reference corpus.

Lang-8 Learner Corpora mainly includes Japanese L1 learners of English essays followed by Mandrin, Korean, and otherL1 backgrounds including Arabic (eight in total).Itis comprised of 580,549 words including the reference native speakers' corpus. However, Arabic-speaking learner data comprises a fraction of this corpus (737 words only). According to Brooke and Hirst (2013),the corpus is disproportionally represented. Arabic constitutes only 0.19% of the whole corpus. As was the case with the ETS corpus, this corpus was divided into training data and testing data for research purposes. The corpus can be distributed for research or educational purposes only.

The Montclair Electronic Language Database (MELD) learner corpus is part of a project at Monclair State University in the USA. It is made up of essays written by 18 non-native speakers from various L1 backgrounds. The database is unique in the sense that it includes annotation of learners' errors. It is comprised of 44,477 words of annotated texts and 53,826 of unannotated data so far. The website[4] of this corpus also includes the metadata of the learners including their sex, age, native country, L1, L2, L3, and L4 information. As was the case with Lang-8 Learner Corpora, there is only one Arabic-speaking participant representing Arabic-speaking learners in this corpus which makes it difficult to use it in its own right. Presumably, there are many other Arabic-speaking learner corpora that are not publically available. As has been shown, there is limited number of Arabic L1 learner

[3]Website: https://catalog.ldc.upenn.edu/LDC2014T06
[4]https://www.montclair.edu/chss/linguistics/department-research-projects/meld/talks/montclair-electronic-language-learner-database/

corpora, and most of which are unavailable as access to them is restricted. The available ones can be used limitedly due to what they represent as the BALC and QLC. Others as Lang-8 Learner Corpora are very limited in scope in what they can represent which calls for the need for more Arabic L1 learner corpora.

4. The Present Corpus

4.1 Rationale

The review above clearly shows that the available Arabic-speaking corpora so far clearly lags behind other L1 backgrounds. Consequently, learner corpus research of Arabic-speaking learners is also still in its infancy and more corpora must be compiled. The rationale behind the compilation of COLTLC can be explained in terms of the lack of Arabic-speaking learner corpora that can match the research interest for which COLTLC was compiled. The other reason can be attributed to the fact that Arabic is a language that has developed into a number of different spoken vernaculars across the Middle East region and North Africa. English is spoken as a foreign language in some Arab countries (e.g. in Saudi Arabia) or as a (second language in others, e.g. in the UAE). Therefore, even the available accessible ones cannot be used unless dialects or countries are clearly marked and well proportioned.

A huge source of differences amongst the Arabic learners lies in the fact that some Arab countries have been greatly affected by other languages due to colonisation (e.g. Algeria). Therefore, it goes without saying that eventhough the background is Arabic, the discrepancies in the way people across the Arab nations use the language is clearly marked. Corpus linguistics methods are ideal for research on register and dialect differences and Arabic inter-learners studies is clearly overlooked. Taken together, these factors prove that the interlanguage of the Arabic-speaking learner of English can be exceptionally unique. Compilations of learner corpora, as well as corpus-based research are needed to shed light on the characteristics of the Arabic L1learners across the Arab World. Since that learners have tendencies to make certain errors based on their L1 background (Swan and Smith, 2001), it would be interesting to compare Arabic L1 speakers data from different countries. Many important patterns of production might be unnoticed until they are scrutinised by analysingconcordances of positive/negative transfer-prone items. This can provide the basis not only for raising awareness of the learners about errors they unconsciously make, but also the structures that are similarly used in their L1 (particularly their informal L1). Moreover, such informationcan be used in the classroom for pedagogical purposes, and for producing tailor-made exercises for the specific areas where learners get confused and struggle. Moreover, developed material can be shared amongst practioners and learner corpus analysts to further their understanding of specific learners' features peculiar to Arabs. It is hoped that the compilation of COLTLC would make a small step forward into this direction.

4.2 Design, size and representativeness

The process of compiling a corpus is a major project that involves a number of decision makings. A random collection of data does not qualify a body of assembled texts as a learner corpus. It must be built according to explicit linguistic criteria in order to be used as a sample of the language (Sinclair, 1996). Granger (2012) in highlighting the importance of establishing design criteria, argued that it should also be aligned to a particular research objective. COLTLC was compiled with a specific goal in mind of (a) describing the use of prepositions in relation to interlanguage of translation majors at KSU (b) the identification of preposition use in relation to proficiency levels. A more general objective is also to help in supporting the current literature and body of research on the Arabic-speaking learner corpora. Based on which, dictionary compilers and material designers rely to produce their materials.

Tono (2003) identified certain design considerations for compiling learner corpora, and she also explicitly said that it is naturally the case that design can vary from one project to another. She mentioned the following: the genre of essays compiled, use of references (i.e. access to dictionaries, source texts), and time limits. She also mentioned the importance of mentioning the method of collection (i.e. is the corpus cross-sectional or longitudinal) along with the method of elicitation. Detailed learner-related information and the task setting are very important when collecting learner corpora. In relation to this, she identified: L1 background, L2 proficiency, L2 environment, the level at school, motivation, attitude, age, and cognitive style. The option of whether to include all the information stated above or not is essentially dependent on the research objective. Cognitive style for instance is not required if it is not part of the research query. A measure of proficiency however, is very important for COLTLC as it helps in exploring the research objective.

Analysis of learners from various levels of proficiency once grouped into sub-corpora opens a new window of limitless opportunities to discover typical problems that learners at a particular proficiency level have. Such aspects are very important to be comprehensively considered.

4.3 Size and Representativeness

Balance, representativeness, and homogeneity are ideals which all corpus builders should strive for, nonetheless they are hard to be attained.Representativeness refers to the extent to which a sample includes the full range of variability in a population (Biber, 1993). By design, COLTLC was compiled to be representative of EFL learners majoring in translation at KSU.Corpus representativeness is not only dependent on the purpose for which the corpus is used but also on the specific linguistic features under investigation. Leech (1991) proposed that "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 generalised to the said language variety" (p. 27). To cut it short, McEnery and Hardie (2011) conclude that the measures of representativeness and balance are matters of degree.

The issue of representativeness in corpus design goes hand in hand with the corpus size. Both are crucial elements that must be taken into account. The size of the corpus depends again on the purpose for which it is intended. Corpus compilation can be a time-consuming labor-intensive work which explains why many researchers may opt for using ready-made corpora to carry out their research (e.g.Gabrielatos, 2013). However, this luxury is not always possible, especially with Arabic-speaking learners as was seen in the review, options are rather limited. Since that corpus is usually designed for a particular research objective, it might not be possible to align the available ready-made ones with one's specific research objective. Therefore, it is very difficult to evaluate a corpus as this will have to depend on what was the corpus built for.

Specific indications or cut-off points for size consideration cannot be simply set or identified; they actually might be problematic because whether a corpus is 'large' or not will depend on what it tries to represent, and what research objective it tries to investigate. As a rule of thumb, Bowker and Pearson suggest it should be of "a greater number of texts than you would be able to easily collect and read in printed form" (2002, p.10). McEnery and Wilson (2001) suggest that the size of the corpus needed to explore a research question is dependent on the frequency and distribution of the linguistic features under investigation. Along the same line, Bowker and Pearson (2002) indicate that a corpus should be 'substantial' especially for making statistical claims of frequency. As a result, if the corpus is small, it can only provide a small window on the phenomenon under investigation and hence, the results will only provide a partial picture of the bigger complex issue. On the other hand, a large corpus will provide a more comprehensive view of the phenomenon, and thus will always be superior to a smaller corpus. There is nonetheless value in studying a small corpus, such value is clearly dependent on how well it serves the research objective and captures features under investigation. Leech (1991) concluded that small corpora may contain sufficient examples of frequent linguistic features. This indeed applies to feature like the English articles (definite and indefinite), and some of the prepositions.

Disadvantages of small corpora are the limitations of generalisability, replicability, and extrapolation of the results. The only solution to this issue is to treat corpus-based findings using a small corpus to deal with it with caution*(Xiao, 2010).* A plethora of influential research was carried out using a small self- assembled corpus (e.g. Cameron and Deignan, 2003) which makes assembling COLTLC worthwhile.

4.4. Corpus material & data collection procedures

The material of COLTLC was collected in the **C**ollege **O**f **L**anguages and **T**ranslation at KSU. The 45 participants were from a homogeneous group of female students taking a simultaneous interpreting course. The data collection was implemented over two consecutive semesters, in the academic year 2015/2016.Participants were not given credits to participate, but as an incentive they were told that they can have their proficiency test scores. Each participant was first required to sign a consent form and fill in a learner profile form requiring information about nationality, age, length of English study, perceived proficiency level, level at university, and years of residence in an English-speaking country (if any). Below, Table 1summarises all the type of data collected in COLTLC.

Teble.1 data collection outline of COLTLC

Session	Description	Tasks
Week 1	Learner profile and consent form collection	Participants were required to fill in a learner profile form and a consent form.
Week 2	Placement test	Participants were given the OOPT placement test.
Week 3	Writing session 1	Participants were required to write a descriptive essay. They were given 5 topics to choose from.
Week 4	Speaking session 2	Participants were required to write an argumentative essay. They were given 5 topics to choose from.
Week 5	Writing session 3	Participants were required to write a narrative essay. They were given 5 topics to choose from.

English is spoken as an EFL in Saudi Arabia and the research objective is interested in students' use of prepositions in relation to their proficiency level. All participants were therefore tested for proficiency in the second session. The online Oxford Placement test (OPT) was employed to gauge the participants' level of L2 proficiency to better understand how the use of prepositions can vary according to it. Table 2 below presents a breakdown of the participants' level of proficiency. Following the OPT guidelines, scores in OPT were converted into approximately equivalent bands to the Common European Framework of Reference for Languages(CEFR). As can be seen in Table 2 below, the level of participants is quite varied ranging from A1 to C1 levels. Consequently, a distinctive feature of COLTLCis that it can later be divided into sub-corpora according to these proficiency bands.

Table. 2 Breakdown of participants' L2 proficiency

CEFER Band	No. of participants	Grouping
A2	5	Lower-level
B1	24	intermediate
B2	10	intermediate
C1	5	Upper intermediate
C2	1	Advanced
Total	45	Advanced

All sessions for essay writing took place at a computer lab in which participants filled the forms electronically, and also typed in their essays over three weeks. All essay writing sessions were timed. Participants were given 60 minutes for each writing session. There were three writing sessions, each of which involving submitting a 300-500 words essay from three genres (descriptive, argumentative, and narrative essays respectively). For each genre, participants were given five topics to choose from. They used WordPad with no spell-check and no access to the internet or other references.

Another decision involved in the compilation of a learner corpus data concerns deciding on how to annotate, and error-tag the assembled rawtexts. While a raw learner corpus is in itself a useful resource, an annotated Part Of Speech POS-tagged learner corpus would have an added value, especially for investigating grammatical categories and faster processing of errors in the usage of prepositions by the learners. COLTLC corpus was tagged using CLAWS POS-tagger, (then manually checked over the prepositions' tags in particularfor precision). As for errors-annotation, Rayson and Baron (2011) suggest that error tagging of learner corpus is time consuming but it can be a stumbling block in the analysis of the corpora. There are a number of ways to go about annotating learner corpus to handle the non-native type of data with its faults and inaccuracies. Granger (2003) for example, labels different types of errors with special tags, and COLTLC will follow this system to tag errors related to the use of prepositions only.

4.5. Corpus Descriptive

The data of COLTLC was collected in an electronic format. All the files were stored in plain text files so that they could be retrieved and used with multiple soft wares. Word Smith which empowers researchers to explore the corpus through a number of concord ancing and word list functions was used for generating preliminary descriptives. Table 3 below presents the basic frequency counts and statistics of the corpus. COLTLC consists of 51,764 word tokens and 5,385 word types.

The token/type ratio (TTR= 10.43), and the Standardized TTR which calculates the running average of TTR in every 1,000 words in the corpus (S-TTR= 40.05). STTR and word length can be good measures of lexical richness, i.e. complexity/simplicity and variation in the text.

Table.3 Basic statistics of COLTLC

Token (running words) in texts	Types (distinct words)	TTR (Token/type ratio)	STTR (Standardised TTR) (basis 1,000 word)	Mean word length (in characters)
51,764	5,385	10.43	40.05	4.29

From the very outset of this paper, it was suggested the COLT LC will be used to explore the use of prepositions by translation majors at KSU. The first step to begin exploring this notion was to shortlist prepositions that appeared in COLTLC. Table4 below shows examples of the most frequently used prepositions in the corpus and it is obvious that (*to*, *of*, and *in*respectively) take the lead. The next step will be to divide the corpus into sub-corpora and analyse how prepositions are used or underused.

Table. 4 Descriptive of the most frequent prepositions in COLTLC

No.	Preposition	Frequency hits	Percentage %
1.	TO	1713	3.30
2.	OF	909	1.76
3.	IN	900	1.74
4.	FOR	472	0.91
5.	WITH	369	0.71
6.	AT	258	0.50
7.	ON	245	0.49
8.	FROM	199	0.38

5. Conclusion

At the very outset, this paper argued in favour of learner corpus-based approaches for SLA studies to yield strong, evidence-based findings. SLA research has largely favoured experimental and introspective data and dismissed natural language use data. Combining SLA methods and corpus-based methods will hopefully provide a rich type of data which can be very informative for SLA research. The paper has also reviewed the available Arabic L1 learner corpora and, argued for the need to compile a corpus KSU to serve the objective of exploring the use of prepositions by EFL learners. This paper has also shed light on the need for well-constructed and well-proportioned learner corpora for SLA research. While abundant learner corpora are available now, the Arab region in particular lags behind other widely spoken languages. The available Arabic-speaking learner corpora are very limited in scope, number, and variation. There is a need for varied learner corpora of Arabic-speaking learners covering a wide scope of variables to obtain a fine-tuned perspective about the features and patterns of learning of the Arabic speakers in relations to other L2 learners. The rationale and the decisions involved in the design and process of compiling COLTLC were also discussed, and the raw descriptive and statistics of the corpus were presented with the figures of the most frequently used prepositions. The next step is to use COLTLC for in-depth investigation.

References

Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.), *Using corpora for language research:* Studies in the honour of Geoffrey Leech (pp. 248-259). London: Longman.

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. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8(4), (pp. 243-257).

Blanchard, D. Tetreault,J. Higgins,D. Cahill, A.&Chodorow, M. (2013). TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service.

Brooke, J. &Hirst, G. (2013).Native language detection with 'cheap' learner corpora. In S. Granger, G. Gilquin& F. Meunier (eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, (pp.37-47).

Cameron, L. &Deignan, A. (2003). Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol*, 18(3) (pp. 149–160).

Crystal, D. (2003). *English as a global language*. Ernst KlettSprachen.

Ellis, R. (1994). *The Study of Second Language Acquisition* (Oxford: Oxford University Press).

Farr, F. (2008).Evaluating the use of corpus-based instruction in a language teacher education context: perspectives from the users', *Language Awareness* 17(1) (pp. 25-43).

Gabrielatos, C. (2013). *If*-conditionals in ICLE and the BNC. In S. Granger, G. Gilquin& F. Meunier (eds), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use, Presses Universitaires de Louvain, (pp.155-166).

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: Longman.

Granger, S. (2002).A bird's-eye view of 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: John Benjamins.

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

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.

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system.In S. Granger, J. Hung & S. Petch-Tyson.(Eds.), *Computer Learner Corpora, Second Language Acquisition and ForeignLanguage Teaching* (pp. 77-118). Amsterdam: John Benjamins.

Lee, D. Y. W. (2010). What corpora are available? In McCarthy, M. & O'Keeffe, A. (Eds.), *TheRoutledge Handbook of Corpus Linguistics* (pp. 107-121). London: Routledge.

Leech, G. (1991). The state of the art in corpus linguistics.In K. Aijmer& B. Altenberg (Eds.), *English corpus linguistics.Sudies in honour of Jan Svartvik*. (pp. 8-29). London: Longman.

Leech, G. (1992). Corpora and theories of linguistic performance, In J. Svartvik (Ed.), *Directions in corpus linguistics*, Berlin: Mouton De Gruyter, (p. 105-122).

Leech, G. (1998). *Learner corpora: what they are and what can be done with them*. In Learner English on Computer. London: Longman, xiv-xx.

Lozano, C. &Mendikoetxea, A. (2013). Learner corpora and Second Language Acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier& P. Thompson (Eds.), Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam: John Benjamins.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd. ed.). Edinburgh: Edinburgh University Press. (Original work published 1996).

McEnery, T., &Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Mindt, D. (1996). English corpus linguistics and the foreign language teaching syllabus. In J. Thomas & M. Short (Eds.), Using Corpora for language research (pp. 232–247). Harlow: Longman

Myles, F (2015). Second language acquisition theory and learner corpus research. In: UNSPECIFIED, (ed.) *The Cambridge Handbook of Learner Corpus Research*. (pp.309-332) UNSPECIFIED.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Randall, M. & Groom, N. (2009). Introducing the BUiD Arab Learner Corpus: a resource for studying the acquisition of L2 English spelling.in M. Mahlberg, V. González Díaz and C. Smith (Eds.) *Proceedings of the Corpus Linguistics Conference CL2009*. University of Liverpool, UK.

Rayson, P. & Baron, A. (2011). Automatic error tagging of spelling mistakes in learner corpora. In Meunier, F., De Cock S., Gilquin, G. &Paquot, M. (Eds.) *A Taste for Corpora*. In honour of Sylviane Granger, Studies in Corpus Linguistics, 45. John Benjamins, Amsterdam.

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

Sinclair, J. (1991). *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Sinclair, J. (Ed.). (2004). *How to Use Corpora in Language Teaching*. Amsterdam: Benjamins

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

Sripicharn, P. (2010). How can we prepare learners for using language corpora? In A.O'Keeffe&M.McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp.371-384). Abingdon/New York: Routledge.

Tono, Y. (1999). Using Learner Corpora in ELT and SLA Research. Paper presented at the Symposium on the Roles of Corpora in Language Teaching and Language Engineering of the 12th World Congress of Applied Linguistics (AILA), Tokyo, Japan.

Tono, Y. (2003). Learner corpora: Design, development and applications. In *Proceedings of the 2003 Corpus Linguistics Conference*, D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.), (pp.800-809). UCREL: Lancaster University

Zhao, L.X. (2014) Ultimate attainment of anaphora resolution in L2 Chinese. Second Language Research, 30 (3). (pp. 381-407).

Zhao, Y., &MacWhinney, B. (2010). Competing cues: A corpus-based study of English tense-aspect acquisition. BUCLD Proceedings, 34, (pp.503-514).

Xiao, R. (*2010*). Corpus creation. In N. Indurkhya& F. Damerau (Eds.), The *handbook of natural language processing* (2nd ed., pp. 147-165). London: CRC press.