# Towards building a corpus for Kurdish EFL learners

**Jamal Anwar Taha1**
**University of Sulaimani**

**Assist. Prof. Dr. Hoshang Farooq Jawad2**
**University of Sulaimani**

**Abstract**

**This paper is taken from a PhD dissertation to shed light on the mechanism of building a corpus for Kurdish EFL leaners. First, it attempts to give an overview of the existence of current corpora in different languages then talks about the importance of having a corpus in any language. Since Kurds go through trouble times that is why this academic aspect has not been served that much, thus, this paper attempts to fill this gap and tries best to build a corpus for Kurdish EFL learners. Overall, this paper argues the importance of having corpora in general and a corpus for Kurdish EFL learners in particular to conduct academic research and serve academics in the best way possible.**

**Keywords: Kurdish EFL students, corpus, PhD dissertation, academic research, learner corpus**

**الملخص**

هذا البحث مأخوذة من رسالة الدكتوراه لإلقاء الضوء على آلية بناء المتن لطلاب الكرد يدرسون اللغة الانجليزية. أولاً:هذا البحث يحاول تقديم لمحة عامة عن وجود المتن اللغوي الحالي في لغات مختلفة ثم يتحدث عن أهمية وجود المتن من عدة لغات. وبما أن الأكراد يمرون بأوقات عصيبة ، فهذا هو السبب في أن هذا الجانب الأكاديمي لم يخدم هذا القدر ، ولذلك تحاول هذا البحث سد هذه الفجوة ، وتحاول بشكل أفضل المتن اللغوي للمتعلمين الأكراد في اللغة الإنجليزية كلغة أجنبية.

بشكل عام ، تجادل هذا البحث بأهمية وجود المتن اللغوي بشكل عام و المتن اللغوي للمتعلمين الأكراد في اللغة الإنجليزية كلغة أجنبية على وجه الخصوص لإجراء البحوث الأكاديمية وخدمة الأكاديميين بأفضل طريقة ممكنة.

**پوخته**

ئەم توێژینەوەیە لە نامەی دکتۆرا وەرگیراوە بۆ تیشك خستنە سەر میکانزمی بنیاتنانی مەتنێك بۆخوێندکارانی کورد کە لە بەشی ئینگلیزی دە خوێنن. لەسەرەتادا هەوڵدراوە کورتە باسێك لەسەر ئەو مەتانەی لە زمانە جیاوازەکان لە ئێستادا بوونیان هەیە و لەدوایدا باس لە گرنگی هەبوونی مەتن لە هەر زمانێك کراوە، چونکە کورد بە بارودۆخێکی نالەبار تێپەڕیوە و ئەو لایەنە ئەکادیمیە خزمەت نەکراوە، لەبەر ئەوە ئەم توێژینەوەیە هەوڵێکە بۆ پڕکردنەوەی ئەو بۆشاییە ئەکادیمیە و هەوڵێکی جدییە لە بونیاتنانی مەتنێك بۆ خوێندکارانی کورد کە لە بەشی ئینگلیزی دە خوێنن. بەشێوەیەکی گشتی ئەم توێژینەوەیە باس لە هەبوونی مەتن و هەبوونی مەتنێك بۆ خوێندکارانی کورد کە لە بەشی ئینگلیزی دە خوێنن بە شێوازێکی ئەکادیمی بە شێوەیەك کە خزمەت بە ئەکادیمیەکان بکات بە باشترین شێوە.

## Introduction

Corpora are needed because through corpus evidence for any hypothesis will be given and most importantly human intuition is unreliable to verify that it needs evidence to support the argument by the data. Besides corpora provide more exposure to the natural language use and create genuine novel paradigms for language learning. The use of corpus is important for the following reasons:

1. It studies the patterns in a qualitative and quantitative sense.
2. It checks either to approve or refute our intuition and research questions about language and to further explain or discover new language patterns.
3. It tries to look at grammar in authentic contexts.
4. It provides reliable evidence or information than intuition. (Halliday et al., 2004: p.34)
5. Collection of texts complied with the intention to be representative and balanced variety, register and genre as well as analyzed linguistically.
6. It is representative of the area or speakers of a particular language since there is a sizable proportion of collected data.
7. It is produced in a natural communicative setting since it is written naturally in a natural setting, written for authentic communicative purposes and not notifying the participants that is not telling them that it is for research purposes.
8. A corpus is balanced since the proportion of a particular part is represented and reflected the proportion example that makes up this variety and the importance of the variety.

9. Machine readable due to the fact that all corpora are stored in the form of plain Unicode text file that can be a loaded, manipulated and processed platform independently. (Gries, 2017: p.8-11)
10. Corpora give information about how language works.
11. Classroom teachers motivate students to use corpora to observe nuances of language and draw comparisons between languages
12. Translators also make use of corpora to compare the use of apparent equivalents in a language i.e. monolingually and between two different languages i.e. bilingually.
13. Corpora can be used to establish a norm of frequency in sub-fields of linguistics like forensic and clinical linguistics as well as stylistics.
14. Corpora are also used to investigate cultural attitudes expressed via language (Hunston, 2002: p.13-14).
15. It is a tool said it is like the invention of telescope in the history of astronomy which tells what language is like! (Stubb, 1996:p. 231)
16. We need corpora for everything especially quantitative analysis.
17. Corpora serve as a meta-analysis since it is testable, reproducible on many languages.
18. It shows language variation through corpora.
19. It gives a flexible data analysis.
20. It gives interpretation about words and texts through annotation.

If we go back to the history of corpus, we see that it is not a new topic but corpora were collected for the purpose of making dictionaries since old time by collecting the data first-hand from the people for example in Islamic religion scholars more than a thousand years ago collected Holy Quran and sayings of the Prophet of Islam Muhammad (PBUH) known as traditions complied by Islamic scholars after his death up until recent history, in the Christian world, the collection of Bible in the 13 century by Anthony of Padua is another example of corpus in the 18th century another example is that of Samuel Johnson's dictionary in 1755. In the near history, some small, large and huge corpora were collected from 1960s to 2000s. The small corpus was collected from 1960-1980s by Brown University which was approximately 1 million words focusing on word frequency and concordance used for grammar studies. Regarding the large corpus, COBUILD Longman Cambridge between 1980s to 1990s gathered 18 million words which were used in EFL dictionaries, a corpora in other languages and their impact on language description. Later a huge corpus was collected between the 1990s to the 2000s known as BNC (Bank of English) which adopted more than 450 million

words and after that Oxford English corpus surpassed that number by gathering more than 1 billion words comprised of a variety of texts by focusing on statistical accuracy, corpora in many other languages apart from English and a specialist corpora which focused on language learner, business English and transaction English language. The other largest corpora were created and updated from 1990s until now is the Corpus of Contemporary American English known as COCA which contains more than 500 million words. For further information about the types and size of Corpora, see the following table.

| English | No. of Words | Language/Dialect | Time Period |
|---|---|---|---|
| News on the Web (NOW) | 4.77 Billion + | 20 countries/Web | 2010-present |
| Global Web-Based English (GlowWbe) | 1.9 Billion | 20 countries/Web | 2012-2013 |
| Wikipedia Corpus | 1.9 Billion | English | 2014-Present |
| Hansard Corpus (British Parliament) | 1.6 Billion | British | 1803-2005 |
| Compleat Lexical Tutor | 500 Million + | Canadian | 2000s-Present |
| Corpus of Historical American English (COHA) | 400 Million | American | 1810-2009 |
| Corpus of US Supreme Court Opinions | 130 Million | American | 1790s-Present |
| TIME Magazine Corpus | 100 Million | American | 1923-2006 |
| Corpus of American Soap Operas | 100 Million | American | 2001-2012 |
| British National Corpus (BYU-BNC) | 100 Million | British | 1980s-1993 |
| Strathy Corpus (Canada) | 50 Million | Canadian | 1970s-2000s |


e-ISSN (2522-7130) p-ISSN (2410-1036)

| English | No. of Words | Language/Dialect | Time Period |
|---|---|---|---|
| CORE Corpus | 50 Million | Web Registers | Till 2014 |
| **Other Languages** | | | |
| Corpus del Español | 2.1 Billion | Spanish | 1200s-1900s |
| Corpus do Portugês | 1.1 Billion | Portuguese | 1300s-1900s |

Table (1)

Adapted from Davies, 2017 (corpus.byu.edu)

There are many other known corpora having less words than listed above for example corpora for Arab learners but this chapter only deals with the prominent ones in the field of corpus linguistics. For linguistic researching in English language whether applied or theoretical due to our experience, we believe that Compleat Lexical Tutor ,which is created by professor Tom Cobb from the University of Quebec, is one of the best one because of the following reasons:

1. It analyzes the large amount of text.
2. It produces data such as frequency of words in academic, non-academic, spoken or written registers but mainly concentrates on academic vocabulary that is why it is very useful for practical research.
3. It produces concordance lines for qualitative data analysis.
4. It tells us more about grammar and lexico-grammar in English language.
5. It is constantly updated by specialists in corpus linguistics.
6. It is a useful tool for research in data-driven language learning and language analysis since the data there are reliable and peer-reviewed papers accessible freely to language researchers.
7. It serves ESL teachers worldwide more than 11000 pages in a day.
8. It helps to learn vocabulary in different contexts.
9. It is free of charge and it can be accessed everywhere in the world.

## 2. Types of Corpora

Due to the importance of corpora in researching in all fields of knowledge as they support the claim that they are making in pure and non-pure sciences whether theoretical or applied, there are many types of corpora. Each is used in an area and fits the need of a type of field or sub-field that the researcher explores. This chapter tries to highlight the types of corpora that exist in this vital field then turns the focus to the type of corpus that is going to be studied in this

dissertation. Each corpus is designed for a particular purpose, for example historical corpus is designed to trace the diachronic aspect of a language. Hunston (2002: p.14-16) isolates eight types of corpora, each of which serves a purpose in researching about language.

1. **Specialized Corpus:** this type of corpus is used to investigate a particular type of language or a particular type of text like newspaper editorial, geography textbooks, academic articles in a particular subject, lectures, casual conversation and essay writings by students. That kind of corpus is a text-restricted type to particular time frame in a particular setting for example investigating conversation in a tea-shop, bookshop, newspaper article deals with a particular topic and the language uses by criminals when conducting a crime but there is no limit to the degree of specialization of this kind of text. Examples of that kind are British English specialized corpus which is a five-million word Cambridge and Nottingham Corpus of Discourse in English known as (CANCODE) which is an informal register of English language and American specialized corpus is Michigan Corpus of Academic Spoken English known as (MICASE) comprises of a spoken register in US academic settings.
2. **General Corpus:** this kind of corpus is larger than specialized corpus and more general as its name indicates, it incorporates many types of texts whether spoken or written but it may not be a representative sample. This kind tries to construct a reference material for many purposes like language teaching for teachers, language learning for students or reference material for translators. It may serve as base line to draw analogies with specialized corpus, since it serves as a reference it is called a reference corpus since it aims to provide a comprehensive information about a language and most importantly it attempts to be a representative sample of all varieties of a language so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials. Examples for this kind are British National Corpus known as (BNC) adopting more than 100 million words and Bank of English adopting 400 million words until 2001 consisting of written British English in Lancaster-Oslo/Bergen Corpus (LOB) corpus and its counterpart in the US Brown corpus consisting of written American English which were both compiled in the 1960s comprised one million words in each one of them.
3. **Comparable Corpora:** this type of corpus involves drawing a comparison between two or more corpora collected about two or more languages or varieties with the attempt to pinpoint the differences and similarities between them and also benefit translators to find equivalents in each one of them against the other. This kind of corpus may include the

varieties of a language to identify the dialectal and regional varieties of the same language like collecting corpus about both central and Kirmanji Kurdish, therefore, this kind of corpus works monolingually and bilingually. When constructing corpora about two or more varieties or languages, they should contain the same proportion so as to have a balanced comparison like newspaper texts, novels, lectures, casual conversations and so forth, example for this corpus is the International Corpus of English known as (ICE) which contains more than one million words of different varieties of English like Canadian English, American English, British English, Australian English and so on. For the time being, there is no multilingual corpora except for comparable and parallel corpora to collect words and texts from several different languages.

4. **Parallel Corpora:** this kind of corpus tries to draw analogies between two or more corpora in two or more than two different languages in an attempt to investigate and depict the differences between them. This kind of corpus is especially interesting and a good resource for translators since through these corpora, they find potential equivalent expressions in each language. This kind is especially interesting for a country having two or more official languages in which the government produces a text in two languages simultaneously like that of in Iraq in which Iraqi government produces every instruction in both Arabic and Kurdish languages. Likewise, European Union produces everything in all official languages that are employed in both spoken and written forms in the union synchronously. Because of this, they need parallel corpora to be as much specific as possible in finding the equivalents among the different languages.
5. **Learner Corpus:** this type of corpus tries to focus on the learner and the collection of texts, essays and writings produced by a language learner in the learning process. It aims to pinpoint the aspects that learners differ from one another on one hand the language of a native speaker on the other. Thus, it is like a comparable corpus to identity the gaps that learners encounter in the learning process then identifying these lacunae to be properly dealt with in the best way possible. Therefore, the learners and their obstacles in language learning is under the limelight. Examples for this kind are International Corpus of Learner English known as (ICLE) is the best example adopting a collection of 20000 words through which each essay written by a learner of English from a particular language background like Arabic, Kurdish, French so forth, and Louvain Corpus of Native English Essays known as (LOCNESS) which a comparable corpus of essays written by native speakers of English.

6. **Pedagogic Corpus:** this kind of corpus first used by Willis(1993) aims to have a corpus of all the language that the learners exposed to, this corpus for most learners does not exist in s a physical form. To collect that corpus, researchers try to include course books like lectures, books, tapes as well as collecting word instances occur and learners come across in different contexts in the hope of raising awareness among learners. This kind of corpus can be compared with a corpus of naturally occurring English to check or ensure that the learner is being presented with or exposed to the language that is sounds natural and useful.
7. **Historical or Diachronic Corpus:** this corpus attempts to gather a corpus of texts from different periods of time which aims to trace the development of different aspects of language over time especially how words changes orthographically and semantically. The best known example for this type of corpus is Helsinki Corpus consists of texts from 730 to 1713 comprising of 1.5 million words which was compiled during 1984-1991. *The Helsinki Corpus of English Texts* is a structured multi-genre diachronic corpus includes periodically organized text samples from Old, Middle and Early Modern English.
8. **Monitor Corpus:** this kind of corpus is designed to track the current changes in the language which are added daily, monthly and annually. For this reason this kind of corpus is large in size and it increases rapidly, thus, the text type is constant and can be comparable to the previous years in their text size and the quality of lexemes it espouses. This model gave rise to the idea of *rate of flow* as the best way of managing the corpus. Instead of setting 10 million words as the proper proportion of that genre, the setting could just as easily be 10 million words a year. Or a month, or a week. The language would flow through the machine, so that at any one time there would be a good sample available, comparable to its previous and future states. Such a model opened up new prospects for those interested in natural language processing, and it added another dimension to contemporary corpora -- the diachronic. New words could be identified, and movements in usage could be tracked, perhaps leading to changes in meaning. Long term norms of frequency distribution could be established, and a wide range of other types of information could be derived from such a corpus(Sinclair, 1996 **www.ilc.cnr.it**)
9. **Spoken Corpus:** this kind of corpus should be distinguished from speech corpus, it accounts for informal, impromptu conversations usually stored as sound files with their transcripts. Due to the importance of spoken variety, many language scholars especially interested in that kind of corpus since they believe that it truly represents language since it

is used in most situations and they are impromptu in real time settings (Sinclair, 1996 **www.ilc.cnr.it).**

10. **Synchronic Corpus:** this kind of corpus deals with a variety of a language or a language or it gives a snapshot of a language in a particular point of time, for example, dealing with central Kurdish in the 1960s and collecting words and texts from those years to build a corpus. This corpus is usually static because it can not be modified or added to it.

11. **Annotated Corpora:** unlike raw corpus including the majority of the aforementioned corpora, this type incorporates additional information about a text or lexemes collected. This kind of corpus can be divided into two types: one of them is called lemmatized corpus in which each word in the corpus followed by lemma, a form of under which you would look up in a dictionary. The second one is called part of speech tags corpus in which each word in a corpus followed by an abbreviation giving the word's part of speech. The annotated corpora must be according to the standards of text encoding initiative (TEI) or must be according to the corpus encoding standards (CES) (Gries, 2017).

## 2. Methodology: steps of building a corpus

For building any corpus, there should be some mechanisms or tools of data collection. The tools available for the researchers are one of the following:

1. A questionnaire : whether open or close questionnaire. In the case of this corpus collected from EFL learners, open questionnaire is chosen so as to gather as many words as possible from them.
2. Focus group: this tool is used to get data from a fixed set of participants to gain knowledge of especially speaking and listening by preparing some questions in advance and get them to talk as much as the researcher can.
3. Interview: it is another tool to collect data from EFL learners from a fixed set of participants, after the interview the researcher should transcript the speech into the text and look for the set of patterns as well as analyze them.
4. Survey: is a quick way to gauge the view of respondents in a particular issue and then analyze the speeches to the researcher's utmost.

Corpus-based is used to collect from EFL students, this methodology is more deductive aiming at analyzing the systematic patterns of variation and use by Kurdish EFL students in

an attempt to find the predefined linguistic features like that of modification in their writing skills among them finding proper solutions to their problems regarding the use of modification (Biber, 2012: p. 163).

Since corpus designs tend to be representative, two central points are taken into consideration regarding this paper:

1. In terms of size, the corpus needs to large enough to represent the distribution of linguistic features, in which this dissertation attempts to gather data from at least 400 EFL Kurdish students in four major universities in Iraqi Kurdistan.
2. In terms of composition, the corpus must be deliberately sampled the kind of register under consideration in this case the written register of EFL students in Iraqi Kurdistan of which attempts to investigate the modification problems of EFL students in their written skill(ibid.: p. 162).

An open questionnaire is formed constitutes of 30 items each item aims at a kind of modification and also targets the order of modification which are students asked to answer to know whether they can place them in correct positions as well as their proper usage. Then, the questionnaire is distributed among one hundred junior and senior students in a random manner of those students who are willing to participate in each university and they are asked to answer the questionnaire then return it back without telling them what kind of aspect of linguistic feature is under study so as to ensure the actual and natural data in natural texts which is the aim of this methodology. After that the data is analyzed and interpreted using AntConc 3.5.7 corpus analysis tool for each frequent modification item in a diagram. Finally, the conclusions will be drawn from the findings of the analysis and interpretation of the data collected.

Steps of building a corpus:

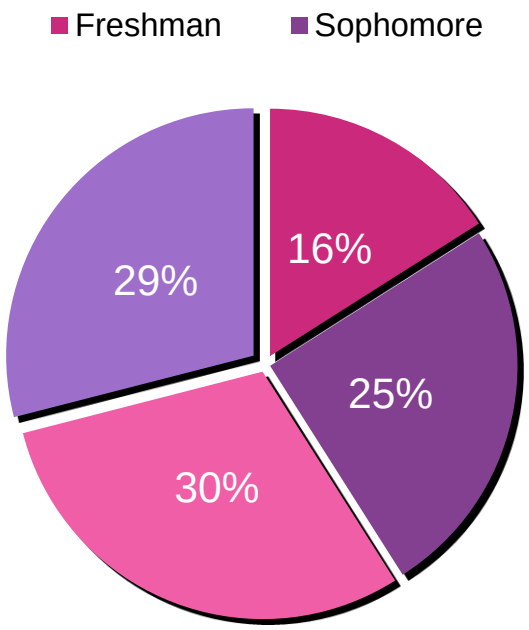
1. First, a kind of corpus should be the target as mentioned earlier in this paper.
2. A suitable tool should be chosen and then a pilot study study should be conducted for the effectiveness of the tool.
3. The mechanism of collecting data i.e. manually or electronically as the learner corpus for this paper is done manually and electronically.
4. There should be a domain to put the date so that researcher can benefit from them.
5. The special kind of corpus software should be chosen to like AntCon, TagAnt, Monoconc Pro, Sketch engine, Word Smith Tool etc. to analyze and process the data .

6. Finally, which aspect of the above mentioned software is widely used i.e. concordance tool, Cluster tool, N-Gram tool etc.

**3. An attempted sample corpus for EFL learners**

In the 400 four hundred students participated in this study, the participants are all university undergraduate students from four universities in Iraqi Kurdistan; Sulaimani University, Salahaddin University, Duhok University and Garmian University, this distribution is a representative of undergraduate program as seen in the following figure.

Figure (1)

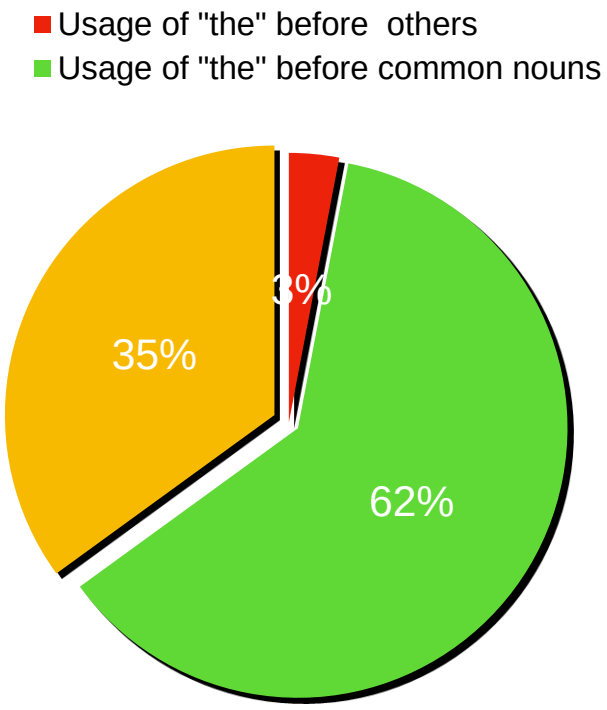



As it is seen in the figure number (1), the majority of junior students participated in the questionnaire which is 30% of them, while 29% of the senior students participated in the questionnaire and answered the questionnaire fully. The percentage drops down to 25% in the sophomore students and the least number participated in the questionnaire was the freshmen students which was only 16% of them partook in the questionnaire. The reason for this is the fact that participating in the survey was voluntary for this end they spent a great deal of time in answering the open questionnaire format designed for this study.

This paper takes the example of "the" in terms of its frequency and its usage in this learner corpus. The definite article "*the*" by far the most frequent one which is 6242 times, this is equal to 2080% two thousand eighty percent in their usages, in only one case the definite article "*the"* is used before the proper noun *Turkey* which is not right. There are 1% one

percent their usages of "the" before verbs like *the feel, the accumulates the produce* etc. there were (3) three cases of using "*the*" before the definite article "the", there were (2) two cases of using "the" with verb

Figure (2)




be like *is, was,* there were (2) two cases of using "the" before predeterminer "all", there were also (5) five cases of using "the" before personal pronouns like possessive pronouns like *the my, the your*. Regarding the use of "the" before proper nouns according to their responses, 8% seven percent of their usages used "the" before proper nouns like *the God, the Japan, the Turkey* etc. through which the definite article is not grammatical before these proper nouns. The use of "the" before common nouns occupy 62% of the total usages of 6242 a pre-modifier before common nouns.

Whereas, 35% thirty five percent of the definite article "the" usages are before adjectives like *the big, the most important, the main, the full* etc. out of these 5% five percent of the 35% thirty percent of "the" usages was before nationality adjectives like *the Kurdish, the English, the American, the Japanese* etc. and 9% nine percent of the definite article's usage was before size

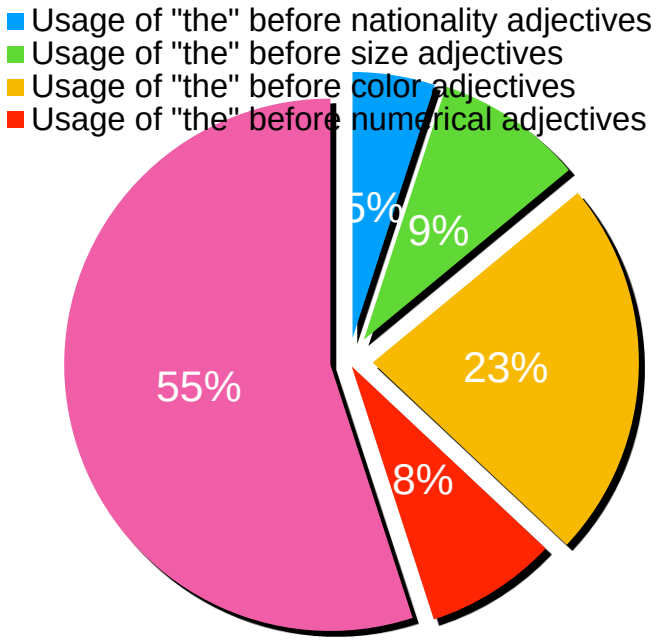



Figure (3)

adjectives like *the big, the biggest, the large* etc. as many as 23% twenty three percent of this definite articles' usages are before color adjectives like *the blue, the green, the grey* etc. and 8% eight percent of the definite article usages is before numerical adjectives and the rest is before the descriptive adjectives.

As far as the usage of definite article with adverbs the respondents only used 1% one percent before adverbs like *the very, the only* etc. and less than 1% one percent they used the definite article "the" before pronouns like *the one(s), the my, the it* etc. and the rest as pointed out in the above with common nouns, adjectives and proper nouns.

The use of definite articles like "the" is misplaced by students a lot, this shows their lack of realization of using this vital determiner in many situations like before predeterminer "all" *the all*

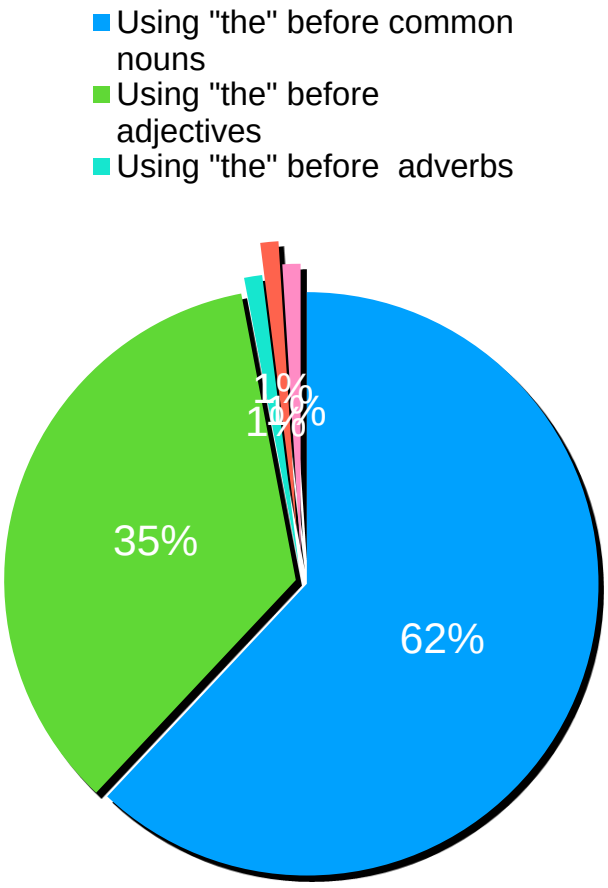



Figure (4)

while in fact it should have been “all the” or using “the” before nouns which are not supposed to be definite or themselves need not be definite like “the Turkey”. The use of this definite article is widely prevalent in their writings as they do not care whether the situation is definite or required a definite article or not.

## 4. Conclusions

The following points can be concluded from the paper:

1. In the absence of a Kurdish EFL students corpus, more should be done to build a corpus for Kurdish EFL learners . This corpus is the starting point and more should be followed.
2. This corpus utilizes concordance tool to pinpoint the modification placement in the corpus and whether Kurdish EFL students have made it right or not.
3. In the light of the analysis “the”, this definite article is by far the most common determiner used by Kurdish EFL students in the four universities chosen from freshmen to senior.

## 5. References

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8.9-37.

Davies, M. (2017). corpus.byu.edu. <Accessed on 20/03/2018>

Gries, Stefan Th.(2017) *Quantitative corpus linguistics with R*. 3rd. ed. London & New York: Routledge, Taylor & Francis Group

Halliday, M.A.K., Teubert, W., Yallop, C. & Cermáková, A. (2004). *Lexicology and corpus linguistics.* London & New York: Continuum.

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

Sinclair (1996). http://www.ilc.cnr.it/EAGLES96/home.html <Accessed on 12/04/2018>

Stubbs, M. (1996) Text and Corpus Analysis Oxford: Blackwell.

## Authors' addresses

**1. Jamal Anwar Taha**
**PhD student in Applied Linguistics**
**Department of English language**
**College of Basic Education**
**University of Sulaimani**
**234 Kaziwa Street- 46001**
**Sulaimani**
**Iraq**
**jamal.taha@univsul.edu.iq**

**2. Dr. Hoshang Farooq Jawad**
**Assistant Professor-Supervisor**
**Department of English language**
**College of Basic Education**
**University of Sulaimani**
**Kurdsat-46001**
**Sulaimani**
**Iraq**
**hoshang.jawad@univsul.edu.iq**