*AMARABAC, Journal of American Arabic Academy for Sciences and Technology*, www.amarabac.com
**Volume 5, Number 12, (2014), PP. 147-166**

# The Use of Corpora in Linguistics

***Reem Ibrahim Rabadi*** [(*)]

**Abstract**: *This article reviews the corpus-based research in linguistics to show the broad use of corpora in linguistics. Corpus input into dictionaries is discussed to reveal applications of corpus linguistics for applied reasons, particularly the amassing of dictionaries. In addition, corpus input in language pedagogy is dealt with to demonstrate the insights of corpora for pedagogical purposes mainly and syllabus improvement in particular. Prior research results for vocabulary input versus output are mentioned which lead to studies of word frequency based on printed materials. This is followed by a discussion of corpus-building, which are: corpus design and collection, analysis of a corpus (analysis of word frequency, collocation, and context), and software programmes used for corpus analysis (WordSmith Tools and other computer programmes).*

**Key words**: *Corpus linguistics, corpus-building, corpus analysis, vocabulary input, word frequency, collocation*

# استخدام الذخائر اللغوية في علم اللغة

***ريم إبراهيم الربضي***

**الملخص**: تستعرض هذه المقالة الأبحاث المبنية على الذخائر اللغوية في مجال اللغويات وذلك بهدف إظهار الاستخدام الأوسع نطاقاً للذخائر اللغوية في علم اللغة. ولهذا الغرض، يناقش الباحث مساهمة الذخائر اللغوية في إثراء المعاجم تجسيداً لأحد نواحي التطبيق العملي لعلم الذخائر اللغوية. وبالإضافة إلى ذلك، يتطرق الباحث إلى مساهمة الذخائر اللغوية في التعليم وذلك بهدف عرض الأفكار المستنيرة التي يمكن الاستفادة منها في هذا المجال لأهداف تعليمية خاصة في بناء مفردات المناهج الدراسية وتحسينها. وتقارن المقالة أيضاً بين النتائج البحثية السابقة حول مساهمة المفردات اللغوية بمخرجاتها وصولاً إلى الدراسات التي أُجريت استناداً إلى المواد المطبوعة حول ما يسمى بتكرار الكلمات. ثم يلي ذلك نقاش حول بناء الذخيرة اللغوية ذاتها بما في ذلك تصميم الذخيرة اللغوية وجمع عناصرها وتحليلها (تحليل تكررا المفردات والمتلازمات اللفظية والسياق والقرينة) والبرمجيات المحوسبة المستخدمة في تحليل الذخائر اللغوية (كأدوات WordSmith Tools وغيرها من البرمجيات).

**الكلمات المفتاحية**: الذخائر اللغوية، تصميم الذخيرة اللغوية ،تحليل الذخائر اللغوية، المفردات اللغوية، تكرار الكلمات، والمتلازمات اللفظية.

(*) German-Jordanian University, School of Languages, reem.rabadi@gju.edu.jo

## *1. Introduction*

Corpus linguistics can be used in many aspects of linguistic inquiry such as syntax, semantics, and lexicography (McEnery and Wilson, 1996). McEnery and Hardie (2012: 228) stated, that "corpus linguistics is ultimately about finding out about the nature and usage of language". Recently, corpus linguistics has been used as a method to investigate natural language usage and as a means for improving classroom materials of language instruction (Reppen and Simpson, 2002; Conard, 2004). A corpus can be a helpful tool to most linguistic investigation according to the purpose of the research. A corpus is generally a sizeable collection of written or spoken resources, or both, comprising natural computer-readable texts to be used in linguistic analysis or to provide new approaches to concepts about some of the general and troublesome topics that require investigation. (McEnery and Hardie, 2012; Baker, 2006; Betty, 2003; Meyer, 2002; Reppen and Simpson, 2002; Donely and Reppen, 2001; Leech, 2000; Biber *et al.*, 1998; Stubbs, 1996; Barnbrook, 1996; Francis, 1982).

The characteristics of corpus-based analyses as elucidated by McEnery and Hardie (2012) and Biber*et al.* (1998) are:

1. It is empirical, analyzing the actual patterns of use in natural texts.

2. It utilizes a large and principled collection of natural texts, known as a 'corpus', as the basis for analysis.

3. It makes extensive use of computers for analysis, using both automatic and interactive analysis.

4. It depends on both quantitative and qualitative analytical techniques.

The following section is about the use of corpus in the field of linguistics.

## *2. Corpus-based research in linguistics*

Corpus-based analysis has supplied new insights into several areas of language structure and usage by making use of large varied corpora in combination with computational and quantitative tools. A collection of approximately 650 corpus-based illustrative works are well documented by the Altenberg (1991) bibliography of corpus linguistics to demonstrate the various usages of corpora in different linguistic fields.

A summary of the different types of corpus-based research in numerous linguistic fields is highlighted in the following paragraphs.

McEnery and Wilson (1996) demonstrated corpus-based studies predating the 1950s in different areas of linguistics with the purpose of demonstrating that there has always been a broad use of corpora in linguistics. These fields of linguistics including McEnery and Wilson (1996) give:

1. Language acquisition: Preyer (1889) and Stern (1924) are examples of linguists whose studies were based on corpora of child language that was collected (1876-1926) from parental diaries recording their child's speech.

2. Spelling conventions: A German corpus of about 11 million words was used by Käding (1897) to collect frequency distributions of letters and classification of letters in German.
3. Language pedagogy: Corpus-based studies of foreign language pedagogy were

used by Fries and Traver (1940) and Bongers (1947).

4. Comparative linguistics: Eaton (1940) compared the frequency of word

meanings in Dutch, French, German, and Italian.

5. Syntax and semantics: Some linguists**,** interested in monolingual depiction,

used the semantic frequency lists used by Eaton. As far as syntax is concerned,

Fries (1952) conducted a study based on a corpus to investigate English

descriptive grammar.

Additionally, McEnery and Wilson (1996) illustrated corpora's participation in several areas of linguistics, including speech research, lexical studies, syntactic studies, semantics, pragmatics and discourse analysis, sociolinguistics, stylistics, historical linguistics, teaching languages and linguistics, dialectology, psycholinguistics, cultural studies, and social psychology.

The use of corpora in different linguistic fields created different kinds of corpora. These will be listed with some examples according to Aston and Burnard (1998) with their commentary. These will be followed by my comments.

1. Geographical types: The Brown corpus of written American English and the LOB corpus of British English.

2. Spoken language corpora: The London-Lund Corpus is the best known spoken language corpus.

3. Mixed corpora: The BNC Corpus is a combination of written and spoken corpora.

4. Historical varieties: The Helsinki corpus of English texts that refers to Old, Middle and Modern English with samples of different dialects.

5. Child and learner varieties: The International Corpus of Learner English is a corpus focusing on learner variety, The Polytechnic of Wales corpus of child language is an example of a child corpus.

6. Genre and topic-specific corpora: The Hong Kong corpus of computer science texts to provide for analysis of technical vocabulary.

7. Multilingual corpora: Several multilingual corpora including texts in both English and one or more other languages have been amassed. An example of this kind is the European Corpus Initiative that involves mainly the most important European languages.

Additional points worth mentioning concern the spoken language corpora that Aston and Burnard (1998) discussed. First, the British National Corpus (BNC) is one of the well-known spoken language corpus today due to its enormous size, with 10% of its overall size (100 million words) representing spoken data. Second, Aston and Burnard did not mention the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) which is another spoken corpus that is well known in this field due to its size (5 million words) of daily British English spoken text (McEnery and Hardie, 2012; McCarthy and Carter, 1997). Third, the London-Lund Corpus cannot be considered nowadays as the best known spoken language corpus as Aston and Burnard suggested because there are other corpora that have appeared, as the above mentioned two corpora.

Another exploration of corpus-based research used in different linguistic fields is by Meyer (2002). The main early reference works to make use of corpora in grammar were *A Grammar of Contemporary English* (1972) and *A Comprehensive Grammar of the English Language* (1985) written by Quirk, Greenbaum, Leech, and Svartvik (Teubert and Čermáková, 2008). Discussions of grammatical interpretation were enlightened by analyses of the London Corpus in several parts of these grammars. Biber *et al.* (1994) stated the importance of corpus-based research in testing grammatical structures that are commonly used for pedagogical purposes. He used a corpus of grammar textbooks to be compared to prototypes detected in naturally occurring discourse so the pedagogical practices would be reconsidered according to the analyses of the study.

Turning now to historical linguistics, corpora provide a reference of knowledge for the linguistic development of English from the past to the present. For example, several regional English dialects and different genres are represented in the texts of the Helsinki Corpus. Another example is the ARCHER corpus (A Representative Corpus of English Historical Registers) that covers the years 1650-1990 of 1.7 million words of American and British English which refer to different types of genres (Lindquist, 2009; Meyer, 2002).

In addition, corpora can make contrastive analyses of English and other languages easy, expand developments in translation theory, and improve foreign language teaching. For instance, The English-Norwegian Parallel Corpus consists of examples of both English and Norwegian fiction and non-fiction. In addition to what has already been mentioned, corpora are a helpful resource to study first and second language acquisition. For example, the CHILDES (Child Language Data Exchange) corpus has created a remarkable quantity of research on language acquisition extending from studies of children learning Germanic and Romance languages to studies of children with language disabilities. One of the largest learner corpora devoted to the speech and writings of persons learning English as a second or foreign language is the International Corpus of Learner English (ICLE). Such learner corpora help in studies of contrastive interlanguage analysis (Meyer, 2002).

After going through a summary of the different sorts of corpus-based research, the following sections will focus on the specific application of corpora in the lexical field. This will include input into dictionaries, frequency lists, and pedagogical applications.

### *2.1. Corpus input into dictionaries*

Some of the first applications of corpus linguistics were for applied reasons, particularly the amassing of dictionaries (Biber and Conrad, 2001), as with the corpus-based dictionary the Collins *COBUILD* Dictionary, 1987 (COBUILD stands for the Collins Birmingham University International Lexical Database) which produced a number of dictionaries derived from the Birmingham Corpus and the Bank of English Corpus (BoE) (McEnery and Hardie, 2012; Lindquist, 2009). In addition, the *Longman Dictionary of American English* was compiled on the basis of a sizeable corpus of spoken and written American English (Meyer, 2002).

When writers of learners' dictionaries select a 'headword list' for dictionary entries, they are actually lead by data from native speakers who speak only one language, and then mainly frequency. However, they may as well involve all the items in familiar textbooks, set reading texts, etc. (Scholfield, 1997). Corpus input is the source of the data that writers of learners' dictionaries need, as Schmitt (2000: 81) indicated, that "corpus input has been particularly useful about the meaning, register, and collocation elements of dictionary entries." Nowadays, the four major English learners' dictionaries: *Cambridge International Dictionary of English*, the *Collins* COBUILD *English Dictionary*, the *Longman Dictionary of Contemporary English*, and the *Oxford Advanced Learner's Dictionary* (Schmitt, 2000) utilize electronic corpora of native speaker English. For example, Longman and Oxford Dictionaries used the British National Corpus (100 million words) to compile the dictionaries (Scholfield, 1997).

### *2.2. Corpus input into word frequency*

Word frequency is another corpus input in the lexical field which shows the number of occurrences of each word in the corpus. Specific computer programmes count word frequency in a corpus whether the words are in base, inflected, or derivative forms. Word frequency can be generated by concordancing programmes (McEnery and Hardie, 2012; Baker, 2006); even an automated frequency count for each single grammatical word can be displayed by using tagged corpus (Biber*et al.*, 1998; Aston and Burnard, 1998).

Listings of general words can be very helpful for learners while learning a language by making frequency counts of large pertinent corpora or even the creation of vocabulary lists for specific purposes, such as the Academic Word List (Coxhead, 2000). Nation and Meara (2002) referred to Michael West's list *A General Service List of English Words*, which includes 2,000 high-frequency words, the "classic list" of the core of useful English words. This example of general word frequency lists was derived from a 5-million-word general written corpus (Biber, 2006). They also indicated that both teachers and learners can benefit from the information derived from frequency studies concerning vocabulary. When high-frequency words are distinguished from low-frequency words, both teachers and learners can usefully spend substantial time making sure that high-frequency words are well learned because these words have to be the major vocabulary target of learners. Learners also need to keep learning low-frequency words after mastering high-frequency words. The needs of learners determine if the number of high-frequency words should be augmented or not.

Word frequency has supplied beneficial insights into the way the vocabulary of English functions. According to Schmitt (2000) word frequency has supplied researchers with three major insights concerning vocabulary. The first insight concerns high-frequency words, which are widely used, and therefore it is very important for learners to learn these words in order to use language properly and to help them to guess correctly the meanings of the interspersed low-frequency words, many of which are necessarily expected to be unknown. Additionally, learners have to learn more than 2,000 meaning senses of the high-frequency words if they want to acquire this essential vocabulary due to the fact that these words are mainly polysemous words.

The second insight is that the majority of the frequent words in English are likely to be *grammatical/function* words that are used only as part of a language grammatical system such as pronouns, determiners, conjunctions, prepositions, and auxiliary verbs. Despite the fact that these grammatical words hold little or no meaning, they are necessary to the structure of English apart from the topic, which is the opposite of *content* words that hold lexical meaning and are affected by the topic of the discourse. The third insight is that spoken and written discourse varies greatly in terms of the use of content words that are used more often in spoken discourse than written discourse.

Corpora have provided other insights as well. The use of interpersonal markers such as (*I think, you know*), single-word organisational markers (*well, right*), apologies, smooth-overs (*never mind*), hedges (*kind of/ sort of*), and a selection of other kinds not likely to take place in written discourse (McCarthy and Carter, 1997). Furthermore, identical words may carry unrelated meanings. For example, the word *got* in the CANCODE Corpus is utilised mostly as *have got* to refer to the possessive verb in its simple form or personal link with something. Schmitt (2000: 74) adds that ''the most frequent fifty words cover a greater proportion of the tokens in spoken discourse than in written.'' A reduced diversity of individual words is normally needed in the spoken discourse. A learner is able to use the language properly in daily conversation with a vocabulary of 3,000 words (Adolphs and Schmitt, 2003) in order to be able to read an average text. Then a person requires more words to perform effectively in the written mode than in the spoken mode.

The frequency of each word form is counted and listed in descending or ascending order of frequency, in alphabetical wordlist order, or lexical bundles (clusters in Wordsmith Tools) (Scott and Tribble, 2006). These obtainable word lists from corpora have a role to play in vocabulary teaching and test development. Analysis of this kind informs researchers how often various words are used, permitting them to distinguish mainly common and uncommon words. This information can be particularly useful in designing teaching materials for language students (Lindquist, 2009; Reppen and Simpson, 2002; Biber *et al.*, 1998; Butler, 1998; Sinclair, 1991).

In addition to computers classifying word lists in alphabetical order, computers are able to list words alphabetically when they are reversed, that is, starting from the last letter of a word rather than the first. This is useful for studies of rhyme, word structure, and so on, or in ascending or descending frequency (Butler, 1985).

### *2.3. Corpus input in language pedagogy*

Corpora provide significant help in language pedagogy in developing teaching strategies for learners of English as a second or foreign language. Carter and McCarthy (2001) referred to the extremely important 'qualitative and quantitative criteria' of the corpus, especially if its insights are related to the language classroom. This can be obtained when a corpus is amassed, analysed and evaluated by using the proper analytical and statistical computer programmes. A notable issue that should be taken into account is that the corpus should enlighten the pedagogic progression, not be operated or dominated by it.

One of the insights of corpus for pedagogical purposes is that learners can gain access to a corpus to use it in order to learn English. They can study genuine examples of language use instead of the fixed examples that are frequently presented in grammar books by inspecting a corpus of native speaker speech or writing with a concordance programme (Granath, 2009). A good example of this is the 'Grammar Safari' methodology that was developed by the Lingua Centre with the Division of English as an International Language at the University of Illinois. Students utilize research engine to obtain grammatical structures in numerous online resources such as magazines, newspapers, and novels (Meyer, 2002). In addition, a number of corpus-based on-line grammars are obtainable from the World Wide Web as the *Chemnitz Internet Grammar of English* (Mukherjee, 2006).In cases where there are not enough computer facilities for students to use, teachers in the classroom could use copies of parts of a corpus or outcomes from corpus explorations (Reppen and Simpson, 2002).

Another insight of corpora for pedagogical purposes as indicated by Donely and Reppen (2001) is to supply teachers with a clear and practical approach for recognizing effective academic vocabulary. They suggested that corpus tools could be used by teachers to improve and amend teaching materials, or to assist students to look into academic vocabulary in the course of classroom activities. Teachers can benefit from corpus tools in creating and adapting materials by assessment, text enhancement and material development. Frequency lists can be used to check vocabulary pre-tests with the purpose of knowing which words students already knew and which words they require to be taught. As for text enhancement, it can be achieved by modifying course texts to emphasise certain words during the course or the academic year. Finally, materials development can be obtained through teachers developing materials to focus on selected words that needed to be taught. This can be achieved by worksheets, listing words in an index, or glossaries.

Corpora can be used as a source for syllabus improvement or classroom resources. The COBUILD Grammar used corpora for syllabi and classroom resources. Owen (1993: 176) referred to these examples of using the corpora for syllabi and classroom resources,

> "they may, for example, attempt to capitalize on increased accuracy in calculating frequency, range, and coverage of lexis for selection purposes (Willis and Willis, 1988). The corpus could be used as a source of strongly collocating elements, or 'lexical phrases' (Nattinger, 1988), which are then incorporated in an exercise. Or, it may be used in a more process oriented fashion, as a direct input to the classroom in the form of concordance pages (Johns, 1988)."

The results of corpus-based research and lists of corpus findings could be used by materials writers. As Conrad (2000) indicated corpus research supplements recommended novelties in grammar pedagogy; the results of the corpus could, for instance, assist a teacher when choosing the issue of consciousness-raising activities.

The use of corpus data in the field of English language teaching has allowed (Aston and Burnard, 1998:19):

1. "more accurate selection of words and senses for inclusion, based on frequency of occurrence;
2. introduction of information concerning the relative frequency of each word and of the different senses of each, and their use in different genres and registers;
3. citation of actual rather than invented examples, selected to illustrate typical uses and collocations".

## *3. Prior research results for vocabulary input vs. output*

It is obvious from the previous corpus inputs that corpus can help in the investigation of lexical inquires that need to be answered. One of these lexical inquires is the effect of vocabulary input on vocabulary output, the purpose of the coming studies. A corpus has to be investigated to examine its word frequency to establish what kinds of words are being used in this corpus. Written corpora, mainly in printed form, are a good resource from which to examine its word frequency as a source of vocabulary input. Therefore, it would be beneficial to look at studies of word frequency based on printed materials.

### *3.1. Studies of word frequency based on printed materials*

Work on frequency word-counts has been performed over recent decades. The most well-known lists will be discussed briefly by way of introduction to the actual studies of word frequency themselves.

*The General Service List of English Words* by Michael West (1953) was the most well-known word-count. West surveyed a corpus of 5 million running words of written English in order to collect the 2,000 most frequently used words in English. The list also contains the frequencies of each words individual meaning senses. (Nation and Waring, 1997; Gairns and Redman, 1986).

*The Teacher's Word Book of 30,000 Words* by Thorndike and Lorge (1944) is a list that was based on a sizeable written corpus of 18,000,000 words. The list has 30,000 lemmas which are based on a count of this old corpus (Nation and Waring, 1997).

The Kučera and Francis list (1967) employed computers in the assembling of a primary list of 2,000 words, which afterwards was developed to 5,000 words (Gairns and Redman, 1986).

*The American Heritage Word Frequency Book* was produced in the 1970s. This list was derived from a corpus of 5 million running words obtained from school

textbooks used over a series of grade levels and over a variety of subject areas taught in schools in the United States. The most useful aspect of the list is that it registers the frequency of each word in the school textbooks for every grade and in every subject of the subject areas (Nation and Waring, 1997).

Ronald Hindmarsh (1980) prepared the *Cambridge English Lexicon* intending it to be adequate for students to pass the Cambridge First Certificate Examination. He listed 4,500 words, with more than 8,000 semantic values. The words were graded on a frequency scale of 1-5, with the scale 6-7 being applied to the less recurrent semantic values of a word. This scaling system makes the list appropriate mainly for course designers (Gairns and Redman, 1986).

*The University World List* was compiled by Xue and Nation in the 1980s. It is based on a wide range of academic texts, and represents 836 high frequency words in an academic list. These words arise in a wide range of academic disciplines, such as maths and science. This list is used by learners for academic purposes (Biber, 2006; Nation and Waring, 1997).

*Word Frequencies in Written and Spoken English* (2001) by Leech *et al.* is the most recent word list that is based on the BNC corpus of 100 million words of both written and spoken contemporary British English. This list provides information about the frequencies of the words in authentic use. It is of benefit for educational needs (Leech *et al.*, 2001).

Turning now to the studies themselves, these will be presented in chronological order according to the date of the study. The first study was published by Kučera and Francis (1967), *The Standard Corpus of Present-Day Edited American English*, that was collected at Brown University during the period 1963-1964 and which contained 1,014,232 words of natural-language text that was analysed for lexical and statistical information. The word frequency of the corpus was listed in descending order of frequency and in alphabetical order. The text of the corpus was synchronic; texts presented in a single calendar year are integrated and representative of a broad variety of styles.

The purpose of the word-frequency lists was not to reveal essential vocabulary or the most common words in English, but to present valuable data for the enhancement and expansion of statistical measures of linguistic analysis in the hope that more suitable ones would be found which would enable this expansion to be carried forward. The most frequent words were function words. Numbers, proper nouns, letters (alphabetical letters), and words with no meaning were included in the lists. This study can be useful in terms of noticing that the most frequent words are function words, in addition to making clear to any researcher that there is a need to proofread the list of words in order to ensure that it is free of any irrelevant words, superfluous numbers or alphabetical letters.

A second study is a French study that was conducted by Lyne (1972) to establish teaching resources or reference works connected exclusively to French business correspondence (FBC) founded on French word-frequency count. Lyne's corpus was made up of 670 originals or consistent copies of French business letters covering the period 1962-67. Three lists of frequency and a concordance were the output of the COCOA programme: the alphabetically- ordered list, the frequency -

ordered list, and the list in decreasing order of 'registral value', and a concordance of all the words except for the highest frequency words that were held back for economic reasons.

It was obvious from the frequency-ordered list that grammatical words, such as *de*, *le*, *la*, *nous*, *vous*, and *être*, were at the top of this list. As for the list in decreasing order of 'registral value', it was created by comparing the frequency of every word ("item" , as Lyne used this term to stand for a main entry) in the FBC corpus to the same word frequency in the Juilland *Frequency Dictionary of French Words*. In order to overcome the differences between the two corpora with regard to their sizes, the frequency of each word in the FBC corpus was multiplied by a suitable factor. The result of this procedure was a new list of the words in the FBC corpus displayed in decreasing order of what one might call the "registral value". The most positive registral items were at the head of this list and the most negative registral items were at the end of this list. Negative registral items according to Lyne (1972: 108), ''the items which occur more frequently and less frequently respectively than one would expect from the Juilland list''. This study indicated that grammar words usually come at the top of frequency words. It demonstrated that such was the case for the French language as has been shown in previous studies for the English language. It also indicated the usage of frequency lists for pedagogical usages.

A third study is the Australian Corpus that Collins and Peters (1988) discuss, which was assembled in 1985 by Pam Peters, David Blair and Peter Collins. One million words composed of samples of 2,000 words were collected from 33 daily newspapers and 11 weekly newspapers investigating 5 reportage subjects: politics, sports, news, finance, and living (society and culture). This corpus was compared to the American Brown corpus and to the British LOB corpus in order to compare American, British and Australian English.

The comparisons between the three corpora were: word frequencies and sets of words frequencies, frequencies of content words and function words, morphological comparison, and orthographical comparison. The researchers concluded that the Australian frequencies of lexis, morphology and orthography sometimes have a tendency to endorse the American approach and sometimes the British. This study indicated that the combination of British and American practices points towards both the past cultural history of the country and its evolving cultural independence. The Australian Corpus is an interesting study which reveals the effect of American English on Australian English. The common belief is that Australian English is mainly affected by British English, but the opposite has been found to be the case, according to this study.

## *4. Issues of building a corpus*

### *4.1. Corpus design and collection*

A corpus, as defined earlier, is a sizeable and principled collection of written or spoken, or both, of natural texts stored in electronic format so as to become in machine-readable form. A corpus needs to be designed and collected before any analysis is carried out.

The design of the corpus is one of the most vital factors in corpus linguistics due to its influence on the whole analysis of the corpus and on the inferences for the reliability of the consequences. It is important that the whole design of the corpus reveals the queries being investigated (McEnery and Hardie, 2012; Baker, 2006; Reppen and Simpson, 2002). For instance, if a researcher is comparing patterns of language found in formal and informal conversations, the corpus should include a collection of possible spoken texts, so that the information originated from the corpus reveals the potential differences in the patterns being compared across the two registers.

When a corpus is created, the collected data has to be converted into electronic versions, stored and organized before any kind of analysis is carried out. A written corpus requires less effort to collect than a spoken corpus because a spoken corpus has to be transcribed into written text, an essential stage in the collection of spoken corpora and is a time consuming process loaded with difficulties (McEnery and Hardie, 2012; Baker, 2006; Reppen and Simpson, 2002; Leech *et al.*, 2001; Leech, 1991). A corpus (Baker, 2006; Reppen and Simpson, 2002; Meyer, 2002; Barnbrook, 1996; Leech and Fligelstone, 1992; Sinclair, 1991) can be machine-readable by converting it into an electronic form with an optical scanner and the OCR (optical character recognition) software operated with it. The OCR decodes the optical image of a page of printed or typewritten data, in a broad range of type sizes and fonts.

The major problem linked to the use of scanners is their inclination to error, so good proof-reading and manual editing is needed to make the scanned texts appropriate for use in research. With the purpose of minimising the opportunity of error throughout the input procedure, error correction is vital for the texts to be prepared to be used in the research. This can be done by using the spell checker, 'find and replace' function within the word processor, or manual checking (Baker, 2006; Barnbrook, 1996).

Two aspects are important in designing a corpus; the size of the corpus and the material chosen to be included in the corpus (Aston and Burnard, 1998; Francis, 1982). Aston and Burnard (1998) and Reppen and Simpson (2002) state that it is generally assumed that the bigger the size of the corpus the more data it provides. On the other hand, Aston and Burnard (1998) caution that in order to recognize certain linguistic trends in a big corpus reliance on automated or partially automated processes will not be available in many fields.

There are general points that any corpus has to comply with according to Barnbrook (1996); these are the following:

1. Contents: the corpus as a sample.

The availability of data should be taken into consideration before a researcher decides precisely what is needed to be examined. Generally, the corpus is a sample of a large collection of language, and is meant to make conclusions available to be obtained about this sizeable body rather than about the corpus itself.

2. The size of the corpus has to be thought of when compared with the requirements of the study. The size of the corpus relies on the central issue of the study. For

example, the BNC corpus has an enormous size of 100 million words to cover the purpose of investigating everyday written and spoken English.

3. Sources of the texts.

There are a number of means of obtaining computer-readable texts. Text archives are one of these means where compilations of computer-readable texts have been collected at large academic sites for several years. The Oxford Text Archive and Project Gutenberg are main sites from which English language texts are presently accessible. Many of the text archives can be obtained by using the internet. However, there is a very serious limitation in the composition of a corpus gained from these sites because it will be controlled by the variety of texts they hold.

Other means include CD-ROMs, although the variety of texts obtainable in this form suffers from similar drawbacks that are found with text archives. Another means is printed or manuscript forms of text, which needs to be converted to computer-readable text before using it.

It is apparent from the aforementioned points that corpus design and compilation can face potential problems. Butler (1998) is one of the researchers who listed these problems. First, using a scanner to convert extremely sizeable amount of texts to computer-readable form would lead to results that are by no means 100 per cent precise, such that proof-reading is required. Second, if the corpus deals with spoken material, then spoken material needs to be collected and transcribed into written texts before conversion to electronic form. Third, it is hard for a researcher to guarantee that the collected corpus in fact represents the type of language the researcher wants to investigate. Other two problems were added by Baker (2006) which are: online texts that need to be formatted and the copyright of the authors and publishers of the work.

A solution for these problems suggested by Butler (1998) is to use an existing corpus on condition that it would supply the researcher with the data sought in the required pattern.

### *4.2. Analysis of a corpus*

#### *4.2.1. Analysis of word frequency*

A word frequency list is produced by identifying each word form found in the text, counting identical forms and then listing them with their frequencies in an alphabetical or frequency order. The lists of word frequency from different corpora or from different parts of the same corpus can be compared to discover some basic lexical differences across registers (Reppen and Simpson, 2002; Barnbrook, 1996).

One of the lexical differences across registers to identify is the number of function/grammatical words and the number of content words in the text. Function/grammatical words are vital to the structure of the text but do not contribute directly or independently to its meaning, while content words are crucial to the meaning of the text. These pieces of information are linked to a knowledge of the nature of the text or to similar pieces of information for other texts. They can present a very beneficial base for an initial study of the text and an examination of its most important features as the pieces of information were gathered from

examining the number of function and content words in the texts of the corpora. Additionally, word frequency lists can supply a useful angle on the essential features of the text. It is also very useful as a basis for choosing words for more information (Barnbrook, 1996).

A useful range of statistics based on word frequency lists are produced by using word frequency software. Total tokens (words) and total types (word forms) are the most well-known totals calculated for word frequency lists. These totals help in showing the degree of lexical variety within a text by showing the distribution of tokens between types in a text. These totals may also supply a basis for examining lexical differences between different types of text (McEnery and Hardie, 2012).

### *4.2.2. Collocation and context analysis*

Collocation is the co-occurrence patterns of words. Collocation shows "the company which individual words keep often helps to define their senses and use." (McEnery and Wilson, 1996: 71). Schmitt (1997: 327) defined collocation as "the syntagmatic relationship between words which co-occur in discourse. Collocations vary in strength from frozen and absolute (as in the idiom *kick the bucket*) through strong and restricted (*blonde hair*) to weak (*nice hat*)." Sinclair (2004:19) described collocation as "the choice of one word conditions the choice of the next, and of the next again".

Collocation analysis usually concentrates on the degree to which the actual pattern of word occurrences are different from the pattern that would have been anticipated. Any significant difference can be taken as a minimum initial confirmation that the existence of one word in the text influences the occurrence of the other one way or another. Sinclair (1991:170) viewed collocation as "the occurrence of two or more words within a short space of each other in a text." Collocational norms within a span of up to four words affect the possibilities of lexical items in English. Collocations can be remarkable and worthy of note because they tend to be unanticiptable, or they can be significant in the lexical structure of the language because of being recurrently frequent (Sinclair, 1991).

As for the kinds of collocations, two main kinds of collocation analysis are used for the texts of a corpus. First, grammatical/syntactic collocations in which a main word as a noun, verb, or adjective 'fits together' with a grammatical word as to be followed by a preposition. For instance *stick to*, *aware of* and *way in*. Second, semantic/lexical collocations that include patterns of two basically 'equal' words such as noun + verb (*wind blows*), verb + noun (*have fun*) and adjective + noun (*happy face*). There is a third kind of collocation that does not belong to either of the previous two. This collocation is for prepositions of time such as *on* Sunday and *at* four o'clock; the reason for using these prepositions in certain situations is still ambiguous (Schmitt, 2000). Collocations are found in the text of a corpus with the aid of concordancers, which are computer programmes used to analyse collocations in a corpus.

Computer programmes are the main tools for corpus analysis. The following section is about these tools.

### *4.3. Software programmes used for corpus analysis*

The information that a researcher will be investigating cannot be obtained from the electronic texts unless certain software programmes are used to search and help in analysing the corpus.

#### *4.3.1. WordSmith Tools and other computer programmes*

Several software programmes are used for corpus analysis, such as the concordance programme, Oxford Concordance Programme (OCP), WordCruncher, and WordSmith. These concordancing programmes permit users to look for particular target words in a corpus, presenting thorough lists for the incidences of a word in context. They make the frequency information, collocation, and other analysis available to the researcher (McEnery and Hardie, 2012; Biber*et al.* 1998). McEnery and Hardie (2012) listed the latest generation of corpus analysis tools, which are the WordSmith Tools (Scott 1996), MonoConc (Barlow 2000), AntConc (Anthony 2005) and Xaira.

In addition, 'generalisable systems' is considered as the latest corpus analysis tools that began as websites. Three of these systems are: the one developed for the BNC by Mark Davies, SketchEngine for lexical and lexicogrammatical patterns, and BNCweb and its clone CQPweb (McEnery and Hardie, 2012).

Meyer (2002) comments that concordancers have statistical capacities built into them to help linguists to analyse specific linguistic structures such as collocations and to check if the outcomes achieved are significant.

## *5. Computer-readable English corpora*

The origins of modern corpus linguistics of the computerized form could be tracked down to the beginning of the 1960s. The first generation of computerized corpora refers back the Brown Corpus compiled by Francis and Kučera that became obtainable for academic research in 1964. It consists of one million words of different types of written American English (Leech and Smith, 2005; Leech *et al.*, 2001; Leech, 1987). Since that time computerized corpora have started to expand its extent and effect. The usefulness of the corpus as a resource of analytically retrievable data has affected its usage broadly to test linguistic hypotheses. Additional important usage of it has been the discovery that the computer corpus presents a new method for creating vital natural processing techniques (Leech, 1991).

There are numerous computer-readable English corpora to discuss and describe in detail (a comprehensive descriptive list for these corpora can be found in Meyer, 2002; Leech, 1991; Taylor *et al.*, 1991). An alphabetical list of 15 different corpora chosen from the above mentioned references. Include,

1. The American National Corpus
2. The Birmingham Corpus
3. The British National Corpus (BNC)
4. The Brown Corpus
5. The Cambridge International Corpus

6. The Corpus of Middle English Prose and Verse
7. The English-Norwegian Parallel Corpus
8. The Helsinki Corpus
9. The International Corpus of Learner English (ICLE)
10. The Lancaster-Oslo-Bergen (LOB) Corpus
11. The London-Lund Corpus
12. The Nijmegen Corpus
13. The Santa Barbara Corpus of Spoken American English
14. The TOSCA Corpus
15. The Warwick Corpus

The BNC Corpus and the London-Lund corpus will be discussed in some detail as a short presentation of a corpus both spoken and written.

The London-Lund corpus of Spoken English (LLC) originates from two projects. The first project, the Survey of English Usage (SEU), began in 1959 at University College London by Randolph Quirk. The second project, the Survey of Spoken English (SSE), was launched in 1975 at Lund University by Jan Svartvik (Lindquist, 2009; Greenbaum and Svartvik, 1990). It was a corpus of one million words containing 200 texts, each of which is composed of 5000 words. Half of these texts were taken from different types of spoken British English and the other half was obtained from various types of written British English. They were collected and analysed to supply the resources for improved descriptions of the grammar of mature educated speakers of English, which was the primary aim of the Survey of English Usage (McEnery and Hardie, 2012; Stubbs, 1996; Greenbaum and Svartvik, 1990). Both dialogue and monologue were included in the spoken English texts. Not just printed and manuscript material texts made up the written English texts, as examples of English for 'spoken delivery' such as news broadcasts, plays and scripted speeches were also included (Greenbaum and Svartvik, 1990).

The spoken texts that were collected and transcribed in London all had to be set in machine-readable form and this was the goal of the Survey of Spoken English at Lund University. The data was put in a condensed transcript and was limited to grammatical analysis (McEnery and Hardie, 2012). As a result of this, the London-Lund Corpus consists of 100 spoken texts (orthographic transcribed with prosodic analysis) that have been distributed, but the 100 written spoken texts have not been distributed (Greenbaum and Svartvik, 1990).

A secondary goal of the London-Lund Corpus was to utilize the computerized data for research. The Corpus's word frequencies were compared to word frequencies of written English in Lancaster-Oslo/Bergen (LOB) British corpus and the Brown University American corpus (Greenbaum and Svartvik, 1990). It is a standard corpus that is available to be used without the need for re-computation (McEnery and Wilson, 1996).

The British National Corpus (BNC) comprises an exemplar collection of some 100 million words which aspires to represent the nature of contemporary (present-day) spoken and written British English in its several social and general usages (Lindquist, 2009; Leech *et al*., 2001; Aston and Burnard, 1998; Butler, 1998). Leech *et al*. (2001: 1) explained what they meant by 'present-day' English:

- " all imaginative texts are dated no earlier than 1960 ( and 80% of them are no earlier than 1975)
- all informative texts are dated no earlier than 1975
- all spoken data are dated no earlier than 1991
- a large majority of BNC texts (over 93%) date from the period 1985-94.''

The BNC contains about 90% written data and 10% spoken data (Leech *et al*., 2001; Aston and Burnard, 1998). The written texts were selected in line with three collection criteria: ''domain (subject field), time (within certain dates) and medium (book, periodical, unpublished texts, etc.)'' (Aston and Burnard, 1998: 29). The domain feature consists of informative texts (factual informative writing) such as correspondence, unpublished reports, reference works, etc.; while imaginative texts consist of different genres of literature, such as poetry, drama, prose, and creative writing. In regard to time feature, informative texts were chosen just starting from 1975 onwards; as for imaginative texts, they were chosen from 1960 and the majority of the texts refer to the period after 1975. Lastly, the medium feature is composed of books, periodicals, miscellaneous published and unpublished materials such as memos and essays. It also includes 'written-to-be-spoken' materials, for instance, play scripts, television scripts, etc. (Baker, 2006; Leech *et al*., 2001; Aston and Burnard, 1998).

Spoken texts form 10 per cent of the BNC's data. This percentage of spoken texts adds up to about 10 million words of transcribed spoken texts (Aston and Burnard, 1998). The spoken texts are divided into two parts; ''a conversational part and a task-oriented part'' (Leech *et al*., 2001: 2). These parts were collected in two different ways (Aston and Burnard, 1998: 31):

- "a *demographic* component of informal encounters recorded by a socially-stratified sample of respondents, selected by age group, sex, social class and geographic region;
- a *context-governed* component (meetings, debates, lectures, radio programmes and the like), categorised by topic and type of interaction."

The subdivisions of the conversational part are: age, social group and sex (Leech *et al*., 2001). Whereas, the subdivisions of the task-oriented part are: educational and informative (lectures, news commentaries, classroom interaction), business (business meetings, trade union talks, consultations), institutional (sermons, political speeches), and leisure (speeches, talks to clubs, club meetings) (Lindquist, 2000; Leech *et al*., 2001; Aston and Burnard, 1998).

The BNC corpus was used by Leech, Rayson and Wilson to produce three frequency lists: alphabetical word order, descending frequency order and statistical uniqueness that shows the difference in frequency for every word in different parts of the corpus (Leech *et al*., 2001).

## 6. Conclusion

One of the purposes of corpus linguistics is pedagogical as indicated by Donely and Reppen (2001). Corpus linguistics provides corpus tools that teachers can benefit from in creating and adapting materials by assessment, text enhancement and material development. One of the corpus tools is frequency lists that both teachers and learners can benefit from, the information being derived from frequency studies concerning vocabulary when high-frequency words are distinguished from low-frequency words in texts. Teachers and learners can benefit from the results of frequency studies, making sure that high-frequency words are well learned as they are the major vocabulary target of learners. Learners also need to keep learning low-frequency words after mastering high-frequency words (Nation and Meara, 2002). Assessment can also make use of frequency lists to pre-test vocabulary knowledge to determine which words students already know and which words they require to be taught (Donely and Reppen, 2001). Another corpus tool is the concordancer that can locate each occurrence of a certain word which can be revealed in any appropriate way.

## References

Adolphs, S., & Schmitt, N. (2003). Lexical Coverage of Spoken Discourse. *Applied Linguistics., 24*(4), 425-438.

Altenberg, B. (1991). A bibliography of publications relating to Englishcomputer corpora. In S. J. a. A.-B. Stenström (Ed.), *English Computer Corpora: selected papers and research guide* (pp. 355-396). Berlin: Mouton de Gruyter.

Aston, G., &Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum International Publishing Group.

Barnbrook, G. (1996). *Language and Computers: A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Betty, K. (2003). *Teaching and Researching Computer-assissted Language Learning*. London: Longman.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins B.V.

Biber, D., & Conrad, S. (2001). Corpus-based research in TESOL Quantitative Corpus-Based Research: Much more than bean counting. *TESOL Quarterly, 35*(2), 331-336.

Biber, D., Conrad, S., &Reppen, R. (1994). Corpus-based Approaches to Issues in Applied Linguistics. *Applied Linguistics, 15*(2), 169-189.

Biber, D., Conrad, S., &Reppen, R. (1998). *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bongers, H. (1947). *The History and Principles of Vocabulary Control*. Worden: Wocopi.

Butler, C. (1985). *Computers in Linguistics*. Oxford: Basil Blackwell Ltd.

Butler, C. (1998). Using computers to study texts. In A. Wray, K. Trott& A. Bloomer (Eds.),

*Projects in Linguistics: A practical guide to researching language* (pp. 213-223). London: Arnold.

Collins, P., & Peters, P. (1988). The Australian corpus project. In M. Kytö, O. Ihalainen& M. Rissanen (Eds.), *Corpus Linguistics, Hard and Soft: Proceedings to the eighth international conference on English language research on computerized corpora* (pp. 103-120). Amsterdam: Rodopi B. V.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*(3), 548-560.

Conrad, S. (2004). Corpus linguistics, language variation, and language teaching. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp.67-85). Amsterdam: John Benjamins B.V.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Donley, K. M., &Reppen, R. (2001). Using Corpus Tools to Highlight Academic Vocabulary in SCLT. *TESOL Quarterly, 10*(3), 7-11.

Eaton, H. (1940). *Semantic Frequency List of English, French, German and Spanish*. Chicago: Chicago University Press.

Francis, W. N. (1982). Problems of Assembling and Computerizing Large Corpora. In S. Johansson (Ed.), *Computer Corpora in English Language Research* (pp. 7-24). Bergen: Norwegian Computing Centre for the Humanities.

Fries, C. (1952). *The Structure of English: An Introduction to the Construction of Sentences*. New York: Harcourt-Brace.

Fries, C., &Traver, A. (1940). *Engllish Word Lists. A Study of their Adaptability and Instruction*. Washington, DC: American Council of Education.

Gairns, R., & Redman, S. (1986). *Working with Words: A guide to teaching and learning vocabulary*. Cambridge: Cambridge University Press.

Granath, S. (2009). Who benefits from learning how to use corpora? In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 47-65). Amsterdam: John Benjamins B.V.

Greenbaum, S., &Svartvik, J. (1990). The London-Lund corpus of spoken English. In J. Svartvik (Ed.), *The London-Lund corpus of Spoken English: Description and research* (pp. 11-59). Lund: Lund University Press.

Käding, J. (1897). *Haufigkeitsworterbuch der deutschenSprache*. Steglitz: privately published.

Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Leech, G. & Smith, N. (2005). Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB. *ICAME Journal 29*, 83-98.

Leech, G. (1987). General introduction. In R. Garside, G. Leech & G. Sampson (Eds.), *The Computational Analysis of English: A corpus based approach* (pp. 1-15). Essex: Longman Group.

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

Leech, G. (2000). Grammar of Spoken English: New outcomes of corpus - oriented research. *Language learning, 50*(4), 675- 724.

Leech, G., & Fligelstone, S. (1992). Computers and corpus analysis. In C. S. Bulter (Ed.), *Computers and Written Texts* (pp. 115-140). Oxford: Blackwell.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow: Pearson Education Limited.

Lindquist, H. (2009). *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press Ltd.

Lyne, A. A. (1972). A Word-Frequency Count of French Business Correspondence: Based on a corpus of approximately 80,000 running words. *IRAL, 10*(2), 95-110.

McCarthy, M., & Carter, M. (1997). Written and spoken vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 20-39). Cambridge: Cambridge University Press.

McCarthy, M., & Carter, R. (2001). Size Isn't Everything: Spoken English, Corpus, and the Classroom. *TESOL Quarterly, 35*(2), 337-340.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Meyer, C. F. (2002). *English Corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: the state of the art- and beyond. In S. Brawn, K. Kohn & J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy* (pp. 5-24). Frankfort: Peter Lang GmbH.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Nation, P., &Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.), *An Introduction to Applied Linguistics* (pp. 53-54). London: Arnold.

Owen, C. (1993). Corpus-Based Grammar and the Heineken Effect: Lexico-grammatical description for language learners. *Applied Linguistics, 14*(2), 167-187.

Preyer, W. (1889). *The Mind of a Child*. New York: Appleton.

Reppen, R., & Simpson, R. (2002). Corpus Linguistics. In N. Schmitt (Ed.), *An Introduction to Applied Linguistics* (pp. 92-111). London: Arnold.

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, N., & McCarthy, M. (1997). Glossary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary Description, Acquisition and Pedagogy* (pp. 327-331). Cambridge: Cambridge University Press.

Scholfield, P. (1997). Vocabulary reference works in foreign language learning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 279-302). Cambridge: Cambridge University Press.

Scott, M. & Tribble, C. (2006). *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins B.V.

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

Sinclair, J. (2004). *Trust the Text: language, corpus and discourse*. London: Routledge.

Stern, W. (1924). *Psychology of Early Childhood up to Six Years of Age*. New York: Holt.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell Publishers.

Taylor, L., Leech, G., &Fligelstone, S. (1991). A survey of English Machine-readable corpora. In S. Johansson & A. Stenström (Eds.), *English Computer Corpora: Selected papers and research guide* (pp. 319-354). Berlin: Mouton de Gruyter.

Teubert, W. & Čermáková, A. (2007). *Corpus Linguistics: A Short introduction*. London: Continuum.