The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus

Mohamed Abdelmageed Mansour, PhD.
Assistant Professor of Linguistics
Department of English
Faculty of Arts
Assiut University
Egypt.

Abstract

It is well-known that Arabic linguistic research in the Arabic countries is not based on corpora because Arab countries do not have Arabic corpus linguistics as compared to the existing English corpus linguistics. Corpora are very important in the advancement of different Arabic linguistics such as sociolinguistics, psycholinguistics, historical linguistics, geolinguistics, contrastive linguistics, grammar, lexicography, stylistics, language pedagogy, and translation. This concise research calls for creating an Arabic National Corpus (ANC) based on four-step design: planning the corpus, collecting the data, computerizing the corpus and analyzing the corpus. This is considered a huge project that needs the collaboration of different national institutions as well as the governments fund and support.

Key Words: Corpus linguistics, corpora, annotation, English corpus linguistics, American National corpus

1. Introduction

In a world of a revolutionary computer technology in the field of linguistics, it seems that the common practice among the Arab linguists in the Arab world is very much frustrating. The only thing that an Arab linguist who is conducting a linguistic research can do is painstakingly sitting in his own office either contriving his linguistic data or extracting his own corpus – a tedious process that involves reading through printed texts and manually recording his data. The linguistic results of this huge effort are not highly accurate because these data are far removed from the real language use, not empirical and lack representation.

The present research calls for creating an Arabic National Corpus (ANC) that will be parallel to the British National Corpus (BNC) and the American National Corpus (ANC) for studying Arabic linguistics empirically rather than introspectively or depending on the linguist's own corpus. To achieve this aim, the idea will be undertaken within the general framework suggested by Meyer (2002) for creating and analyzing linguistic corpora showing how this step-by-step guide can be applied to the linguistic situation in the Arab countries.

2. Corpus Linguistics

Corpus linguistics is simply a tool for linguistic inquiry. That is, it is "a methodological basis for pursuing linguistic research" (Leech 1992: 105). In principle, the corpus, prepared in a computer-readable form, is "a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about language" (Crystal 1991: 89). Corpus linguistics dates back to the sixties, when the first computer corpus, the Brown Corpus, was created. This new trend was not welcomed by generativists who dominated at that time and considered this corpus as "a useless and foolhardy enterprise" (Francis 1992: 28). Corpus linguists, on the other extreme, are not satisfied with the abstract and decontextualized linguistic data supported by generative grammarians, and recently "linguists of various persuasions use corpora in their research and are united in their belief that one's linguistic analysis will benefit from the analysis of real language" (Meyer 2002: 2). The contribution of corpus linguistics in linguistic research is tremendous in many respects. First, the corpus provides the linguist with an empirical data which enables him to form objective rather than subjective linguistic statements.

Second, corpus linguistics helps the researcher to avoid any linguistic generalizations that may be based upon his internalized cognitive perception of language. Third, qualitative and quantitative linguistic research can be conducted in few seconds due to the powerful computers and software that are able to perform complex calculations without errors, hence saving both time and effort. Finally, studying the language empirically can help linguists not only to conduct new linguistic research effectively but also to revise and test the existing theories.

Among the roles of the corpus linguist is to create a corpus; he is a corpus compiler and hence becomes a field linguist because he goes out collecting and recording speech in various locations: homes, offices, schools, universities, work place, and so forth. It is noteworthy that the corpus linguist creates a corpus not for the purpose of studying it. Instead, he creates it for others to study.

Moreover, corpora are used in the field of computational linguistics known as Natural Language Processing (NLP) to develop research in areas such as tagging, parsing, information retrieval and speech recognition systems. These efforts in the field of computational linguistics will help in analyzing the collected raw corpora linguistically. These corpora will provide the linguist with an annotated corpus instead of a lexical or raw corpus.

3. English Corpus Linguistics versus Arabic Corpus Linguistics

There are several English corpora that have been created for the purpose of the empirical study of English linguistics since the 1960s. Undoubtedly, these efforts in the field of corpus linguistics led to the advance of different fields of English linguistics. Some of these corpora are presented here, and for more examples see the Appendix.

-*The British National Corpus (BNC)* contains 100 million words of British English. Ninety percent of the corpus consists of various genres of written English, and ten percent comprises different types of spoken British English (Meyer 2002).

-*Michigan Corpus of Academic Spoken English (MICASE)* is a speech corpus. It was created for the purpose of studying the type of speech used by individuals conversing in an academic setting, such as classroom discussions, students presentations, tutoring sessions, class lectures, and dissertation defenses (Powell and Simpson 2001).
-*A Representative Corpus of English Historical Registers (ARCHER)* is a historical corpus. It covers the period (1650 – 1990) which is divided into fifty-year subgroup texts. Also it is a "multi-purpose general" corpus because it contains many different texts that cover different periods of English (Rissanen 2000).

On the other extreme, although corpora are widely available for English, there is very little available for the Arabic language. In fact, we still have a long way to go before we catch up with English corpora. Throughout the Arab world we do not have one single corpus that we created ourselves and the existed handful Arabic corpora have been created by others who are not Arabs, as shown below:

-The European Language Resources Association (ELRA) provides two Arabic corpora. The first, 140 million words in length, is a corpus of six years work of Al-Nahar newspaper from Lebanon. The second is a corpus of Al-Hayat newspaper and contains 18 million words.

-The Linguistic Data Consortium (LDC), University of Pennsylvania, produced three corpora: a corpus of Arabic newspaper texts containing 76 million words, a corpus of Egyptian Arabic Speech, and a lexicon of Egyptian Arabic. The first corpus is composed of articles from the Agency France Press (AFP) Arabic Newswire. The second corpus consists of 60 unscripted telephone conversations. For each conversation, both the caller and the callee are native speakers of the Egyptian dialect of Arabic who are making calls from inside the USA and Canada. The third one (Gadalla et al. 1998) is a CALLHOME English Arabic corpus of telephone speech and consists of 120 unscripted telephone conversations between native speakers of Egyptian Colloquial Arabic (ECA).

4. Corpus Linguistics and Its Implications in Arabic Linguistic Research

Corpus linguistics serves most areas of linguistics as being the raw material on which the linguistic researcher is working. Corpora are used in linguistic research for the purpose of linguistic description and analysis. The following subsections will explain the role that corpus linguistics plays in different areas of linguistic research with reference to Arabic linguistics.

4.1. Historical Linguistics

There are a number of English historical corpora that contain samples of writing representing earlier periods. These corpora are used to study both language variation in the earlier periods of English as well as language changes and development. For example, the Helsinki Corpus, a 1.5-million-word corpus, contains texts from the Old English through the early Modern English. This corpus has been used by historical linguists to study the evolution of English (Rissanen 1992). Moreover, Skaffari (2009) did significant studies in the middle English words that were borrowed from France. Most of Arabic historical linguistics studies are not corpus based. For example, Wafi (2000), depending on the linguistic data collected manually from different books, studied the history of the Semitic languages: its origin, life and development. His book covered the phonology, grammar, lexicon of these languages, the factors that led to the appearance of different dialects, phonological change and the collapse of some of these languages.

Creating an Arabic historical corpus that represents different periods of Arabic language history will allow historical linguists to investigate systematically the development of, for example, particular grammatical and phonological aspects in the earlier Arabic periods. Moreover, such a corpus can help linguists to study the sociolinguistic variables that affected language usage, such as gender. Various dialect regions can also be studied throughout different historical periods.

4.2. Psycholinguistics

Corpus linguistics serves psycholinguistics. An important corpus that serves the field of psycholinguistics is the CHILDES (Child Language Data Exchange). This corpus contains transcriptions of children learning first and second languages and it has been studied by psycholinguists who are interested in child language acquisition (MacWhinney 2000). It is very important for the development of the study of Arabic psycholinguistics to create an Arabic psycholinguistics corpus. Such a corpus can include, for instance, speeches from Arab normal children who are developing their normal linguistic skills and those who have language disorders such as aphasia and autism. Studying this corpus psycholinguistically may give us a real picture of the normal and abnormal data of the Arabic language of normal as well as linguistically impaired children. Most importantly, forming Arabic psycholinguistic corpus will provide linguists with the chance to conduct contrastive psycholinguistic studies by comparing the linguistic behaviour of Arab children with that of the English ones.

4.3 Sociolinguistics

Corpora can be used to study some sociolinguistic variables such as gender, dialect region, social status and age. For example, in the spoken part of the British National Corpus, Aston and Burnard (1998) used the software program Sara to count the number of instances of the adjective *lovely* spoken by males and females. They found that this word is used more frequently by females than males. The intended Arabic National Corpus can include a sociolinguistic section as an Arabic sociolinguistic corpus. This section can include, for instance, the language of Arab teenagers that can be similar to the COLT corpus (the Bergen Corpus of London Teenage English) that contains the speech of London teenagers (Stenström and Andersen 1996) or the language of educated people. Such a section can help sociolinguists to conduct comparative studies to compare different sociolinguistic variables. It also helps in holding contrastive sociolinguistic studies based on Arabic sociolinguistic corpora and English sociolinguistic ones.

4.4. Lexicography

It is customary that a dictionary provides the users with different kinds of information about words including their meaning, pronunciation, part of speech and examples that give the contextual meaning of the word. Before using the linguistic corpora in lexicography, all this information had to be collected manually and it was time consuming. For example, the *Oxford English Dictionary* took fifty years to complete. The dictionary included five million citations which were "painstakingly collected … subsorted … analyzed" (Landau 1984: 69). Recently, the advancement in computer corpora and software programs changed the way we look at the dictionaries. The use of a software program called the concordancing program that can count the frequency of words in a corpus, detecting affixes and sorting the words by lemmas. As for the parts of speech, if the corpus is tagged, the parts of speech of each word can be automatically determined. Moreover, the lexicographer can also use KWIC (Key Word in Context) format to detect the various contexts in which the word occurs and the meaning of each occurrence.

And if the lexicographer wants to copy the sentence in which a word occurs, it can be automatically extracted from the text and sorted in a file making citation slips sorted in a filing cabinet. Without using the corpus, Fillmore (1992: 39-45) observes that these citation slips that the lexicographer observes will be "largely limited to examples that somebody happened to notice". Thus depending on the corpus, the dictionaries will contain more complete and precise definitions of words since a large number of natural examples from the language are examined.

Not only can the corpus be used to create new dictionaries, but also to revise the existed ones. In this case, the corpus can either supplement or refute the lexicographer's intuitions. To illustrate this point, Atkins and Levin (1995) studied verbs in the semantic category *shake* and quoted its definitions in three dictionaries: *The Longman Dictionary of Contemporary English, The Oxford Advanced Learner's Dictionary* and *The Collins COUBUILD Dictionary*. They found that both the Longman and COBUILD dictionaries list the verbs *quake* and *quiver* as being intransitive, while the Oxford dictionary lists *quake* and *quiver* as being transitive. In calling up all the examples of these verbs in a corpus of 50,000,000 words, they found that both *quiver* and *quake* are used both transitively and intransitively. Thus, the dictionaries have got these verbs wrong.

As far as the creation of Arabic dictionaries is concerned, Al-Eryaan (1984) discussed the stages of collecting Arabic dictionaries manually. The first stage that the Arab lexicographer can do is to gather the words from people living in different regions depending on hearing these words such as 'rain', 'sword', etc. The second stage is to categorize the words under separate headings. The result will be a book for 'rain' and another for 'sword', for instance. The third stage is to gather all this information in a complete dictionary that includes all the words of Arabic. Thus we have Al-Sahah Dictionary, Al-Waseet Dictionary, Al-Kabeer Dictionary, etc. This is the traditional way of creating Arabic dictionaries. The prospective Arabic National Corpus will make the process of creating dictionaries easier, improve the kinds of information contained in them, and address some deficiencies inherent in many of these dictionaries. This can be done by going through a huge number of computerized examples of Arabic that will be included in the prospective corpus.

4.5. Stylistics

If the stylist wants to offer confident stylistic studies, he has to analyze linguistic features of the texts that are computer-readable form. This requires analyzing the literary works to compare between the use of different linguistic devices not only in one's own work but also with other authors' works. This leads to a quantitative analysis of the work – an area where corpora play an important part. Leech and Short (1981) pointed out that stylistics often demands the use of quantification to back up judgments.

Arab stylists who study the stylistic features of the works of some Arabic writers go through their works and write the linguistic features manually – a very tedious and time consuming process. For instance, Al-Trabulsi (1996) analyzed the Anthology "Al-Shawqiyat" written by Ahmad Shawqi, the prince of poets, stylistically. Citing, manually, 11, 320 lines of poetry that cover 370 poems, he studied different linguistic aspects of Shawqi's poetry. However, creating an Arabic National Corpus will help researchers of stylistics to easily and objectively examine the linguistic features of Arabic writers. For example, converting the works of different Arab writers into a computer-readable form will provide the stylistic researchers not only with an effective means of studying the linguistic features of these writers but also comparing their style with that of other foreign writers.

4.6. Pragmatics

Pragmatics means language in context. Corpora are a plentiful source of studying pragmatics. For example, After Stenström (1987) examined what he termed "carry on signals" in a corpus, he was able to classify these signals according to their functions, e.g. *right* has been used in all functions, but especially in a response, to evaluate a previous response or terminate an exchange. *All right* has been used to mark a boundary between two stages in a discourse. *That's right* has been used as an emphasizer. And *it's alright* and *that's alright* have been responses to apologies. Creating an Arabic pragmatic corpus that is a part of the Arabic National Corpus will help linguists to study effectively Arabic pragmatics.

4.7. Contrastive Linguistics

Corpora can be used to facilitate contrastive linguistic analysis. For example, the English-Norwegian Parallel Corpus contains examples of English and Norwegian fiction and non-fiction that are 10,000 – 15,000 words in length. Such a corpus has been used to compare structures in both languages to allow a range of different contrastive studies (Johansson and Ebeling 1996). The appearance of such bilingual corpora led to the invention of the ParaConc program (Barlow 1999) which is used to align sentences in any two languages.

Most, if not all, of the contrastive studies done by Arab linguists are not corpus-based. The unavailability of computerized bilingual corpora has led those linguists to contrive an introspective linguistic data then subject this data to linguistic analysis. For example, Mahmoud (1989) studied the morphological, syntactic and semantic features of middle and inchoative verbs of Arabic and English, Gadalla (1999) gave a morphological and phonological analysis of Standard Arabic and Cairene Arabic, and Mansour (1999) gave a contrastive analysis of the morphosyntax of English and Arabic verbs. These studies are valuable contributions and a step forward in the study of Arabic linguistics. However, had these studies been corpus-based, their findings might have been much different. A Section of the Arabic National Corpus can include a bilingual corpus. That is, collecting and computerizing some foreign texts to be included in the corpus. Creating such a bilingual corpus will facilitate such contrastive studies. Such a corpus might generate an impressive amount of research in the field of contrastive linguistics.

4.8. Language Pedagogy

One of the strategies that can be used in teaching a foreign language is to expose students to extensive training using a corpus. Using a concordancing program to investigate such a corpus will give students real examples of language usage rather than contrived ones that are often found in grammar books. This inductive exploration of different linguistic constructions on vast amount of data will allow students to practice with concordance programs to generate so much data. Creating an Arabic National Corpus will help teachers and students alike to practice Arabic and English grammatical structures by themselves using language in context. Such a process, Gavioli (1997: 84) claims, is an effective "language-learning activity".

4.9. Translation

Bilingual corpora that contain translated texts from two or more languages can facilitate translation studies, train translators and advance linguistic translation theories. Moreover, using such information in translation can be used to create bilingual dictionaries (Schmied and Schäffler 1996). One section of the suggested Arabic National Corpus can include translated works from English to Arabic and vice versa. Gadalla (2003), for instance, studied translating Arabic perfect verbs into English through analyzing manually two Arabic novels by Naguib Mahfouz. A corpus of 250 sentences was randomly and manually chosen from the two novels, 125 sentences from each novel. In fact dealing with such an issue and other similar ones in translation through a computer-readable corpus of a large number of texts may make the task easier and more effective.

4.10. Grammar

Studying grammatical structures, whether in morphology or syntax, yields linguistic information on these structures and their frequency. In the area of qualification, Arts (1992) used the London Corpus to analyze "small clauses" in English. He was able to provide a complete description of the small clauses. In the area of quantification, Collins (1991), in a corpus study, compared the relative frequency of modals in four genres of Australian English: press reportage, conversation, learned prose, and parliamentary debates to test whether modals of necessity and obligation are more suitable for some contexts than others. For more recent corpus study of the history of English syntax, see Rissanen 2012)

In the field of Arabic linguistics, Kebbe (2000), for example, gave a transformational analysis of modern written Arabic based on the transformational theory as formulated by Chomsky. Though the majority of works on the Arabic language concentrated on regional dialects and this book fulfils a long-felt need by focusing on modern written Arabic, it is not corpus-based – a prerequisite that might render the book more realistic. Moreover, Fischer (2000) offered a book which is though unquestionably considered the most useful reference grammar of the classical Arabic language.

However, it is not corpus-based because "the examples cited are for the most part borrowed from the standard grammatical treatises (Wright, Nöldeke, Reckendorf, Bbrokelmann, Wehr, Spitaler) and to a smaller extent are supplemented from my own stock)" Fischer (2000: xiii). In the field of Arabic morphology, Abd-Elghany (1970) studied the morphological units and their role in Arabic word formation, not through a corpus, rather using examples that he obtained either introspectively or cited from other books. Better results might be obtained if his morphological analysis were based on a computer corpus. Creating the Arabic National Corpus will facilitate the study of Arabic grammatical and morphological structures instead of studying them in contrived contexts or depending on the manual corpus gathered by the researcher.

4.11. Geolinguistics

This branch of linguistics studies different linguistic aspects of the languages and dialects in terms of regional distribution. Geolinguistics provides us with maps that present different linguistic features of the different dialects by region. There are different atlases in the world that give dialectical maps for different languages in Europe and America. We do not, as Arabs, have one. The only preliminary effort that was made for Arabs was at the hands of Bergsträsser, the German Orientalist, who applied this idea on the Arabic language for Syria and Palestine (Abd-Eltawab 1997).

In fact creating the Arabic National Corpus can contain a section that includes a survey of the different dialects not only between Arab countries but also in the same country, i.e. creating an atlas for the dialects in Saudi Arabia. . This may help to form the Arabic national geolinguistics atlas which classifies, by region, similarities and differences between different Arabic dialects in phonetics, morphology, syntax and semantics.

5. Designing an Arabic National Corpus

After showing how most of the Arabic linguistic studies cannot dispense with corpora if they are aiming at offering a real picture of a real use of a real language, the next step is how to plan the Arabic National Corpus effectively because as Meyer (2002: 53) states that "well planned corpora are the most effective tools possible for linguistic research". Following Meyer (2002), creating such a corpus goes through four processes: planning the corpus, collecting the data, computerizing the data, and analyzing the data.

5.1. Planning the Corpus

The first thing to consider in planning a corpus is the length of this corpus. Generally, the lengthier the corpus is, the better it will be. The length of the corpus can be determined on two bases: the available resources such as funding and computing facilities. The second basis is the purpose for which it is used, e.g. creating dictionaries and discourse studies need long corpora, whereas grammatical constructions need short corpora. As for the prospective corpus, it is preferable to be a long corpus, as long as it could be, because this corpus is intended to be a multi-purpose corpus. That is, it can cover different features of Arabic language and linguistics.

Not only is the length of the corpus to be considered in the process of planning, but also the length of individual texts. The common practice in corpora compiling is to contain relatively short texts. Biber (1990) found that 1000-word excerpts are lengthy enough to provide valid and reliable information on the distribution of frequently occurring linguistic items. Again the lengthier the text sample is, the better it will be. It is suggested that the Arabic National Corpus contain excerpts that range from 1000- to 2000-words to cover both frequently and infrequently occurring items.

The third thing to be considered by the corpus linguist while planning a corpus is the type of genres to be included in a corpus. The spoken part of the British National Corpus, for example, consists of a variety of spontaneous dialogues, classroom dialogues, monologues, lectures, tutorials, and news commentaries (Crowdy 1993). The written part of the corpus contains various kinds of fictional writing, world affairs, social science and natural sciences. Since the Arabic National Corpus is a huge national project and intended to be multi-purpose; ranging from the studies of vocabulary, to studies of differences between varying regional dialects of Arabic, to grammatical studies, to comparison of linguistic features of various genres of Arabic, it is suggested to include as various written as well as spoken genres as possible.

Some of these genres are: books (e.g. educational, business, natural science, social science, commerce, arts, belief and thought), novels, plays, legal documents, newspapers (e.g. press reportage, press editorials, articles). The spoken part can include the Arabic language in different Arabic countries such as Saudi Arabic, Egyptian Arabic, Syrian Arabic, Tunisian Arabic, etc. The fourth section of planning the National Arabic Corpus is to provide representative written and spoken samples. There are two criteria for including the representative number of texts: linguistic considerations and sampling methodological considerations. Depending on the linguistic considerations, one can choose the most common linguistic genre in the language to be amply represented in the corpus. For instance, in the case of the British National Corpus, spontaneous dialogues are amply represented because all speakers of English engage in this type of conversation (Meyer 2002). Another linguistic consideration mentioned by Biber (1993) is the internal variation of a specific genre. The genre of academic prose, for example, includes subgenres that are linguistically different. They are medicine, the humanities, and the natural and social sciences), hence they should be adequately represented in the corpus.

In gathering demographic sample speech for the British National Corpus "random location procedures" (Crowdly 1993: 259) were used to select the individuals for the samples. Concerning the Arabic National Corpus, it suggested that we use random sampling to gather written and spoken texts. The following is a hierarchal composition of the Arabic National Corpus (adapted from Meyer 2002):

Dialogues: Class lessons, Broadcast discussions, Parliamentary debates, Meeting discussions

Monologues: Speeches, Broadcast news, Broadcast talks, Broadcast interviews, Advertisements

Fiction: prose (novel, short stories, folktales), drama (serious, comic and historical plays) and poetry,

Proverbs: Standard Arabic proverbs and Colloquial proverbs,

Journalistic Arabic: Press reportage – News commentaries – Editorials – Articles – Advertisements – Comments

– Points of view.

Official Documents: Government reports – Business repots – University documents,

Academic Arabic: Class lectures – Class discussions – Students presentations – Tutoring sessions,

Scientific Text: Social sciences, Natural sciences,

Expository text: A summary report about an event – course objectives – critical reports – personal reports – Students examination essays – book reviews.

The fifth step in planning a corpus is controlling the sociolinguistic variables such as gender, age, level of education, and dialect. Gender balance is an important element to observe while planning a corpus in both writing and speech. As for the written texts, the corpus compiler should include a corpus of male and female writers as balanced as possible**.** As for speech, Biber and Burgues (2000) note that in studying the gender variable one needs to consider not just the gender of the individual speaking but the gender of the individual(s) to whom a person is speaking. Thus gender balance can be achieved when the corpus includes a variety of different types of conversations involving men and women as women speaking with other women, men speaking with other men, and women speaking to men.

Texts gathered should be balanced by age. The International Corpus of English, for instance, has the ages 18-25, 26-45, 46-65, and 60-henceforth. This criterion can be applied to the Arabic National Corpus. The educational level is an important variable when planning a corpus because the corpus can be studied according to the educational level. The International Corpus of English (ICE) project, for example, defines the educated persons as those who have at least a high school education level. In planning a corpus, one also needs to include the various dialects that represent regional variation. That is the spoken corpus should be balanced by dialects.

The Arabic National Corpus can be based on the four previous variables. Gender and age balance can be achieved as possible as the corpus compiler can. We can base the Arabic National Corpus on different levels of education. To achieve a balanced corpus on the basis of dialect variation, the Arabic of different Arab countries, and their dialects can be included in the corpus plan.

5.1.2. Collecting Data

Two kinds of samples are collected to be incorporated in the corpus: speech samples and written samples. The first thing that the corpus linguist can do is to collect speech.

5.1.2.1. Collecting Speech Sample

In collecting the speech samples in the British National Corpus, for example, the participants, not the corpus linguist, in the project were given portable tape recorders and instructed to record all the conversations for a period ranging from 2 – 7 days (Crowdy 1993). Digital recorders are sometimes preferred particularly when we work on the computer to edit unwanted background. The individuals record speech in different social contexts such as conversations over dinner, informal conversations among friends, co-workers speaking at work, teachers and students in class discussions, etc. Those individuals can use different of microphones that are suitable for the situation.

According to Meyer (2002) three types of speech are collected: direct speech, telephone conversations, and radio and television broadcasts. Different microphone types can be used with the first type of speech:

1. Uni-directional microphones are used to record single individuals;

2. Omni-directional microphones are used for larger groups;

3. Wireless microphones and laviere microphones (worn around the neck) are used by persons who are moving around and giving speech;

4. Extra-sensitive microphones can be used for recording individuals who are not close to the microphones.

As for recording telephone conversations of individuals talking over the telephone, adaptors can be used to record directly over the telephone. The third source of speech is the radio and television broadcast. In recording this type of speech, one either puts the audio input plug on the tape recorder or connecting the TV to a video cassette recorder by running a line from the audio output plug on the recorder.

In collecting spoken samples for the suggested Arabic National Corpus there are some steps to be followed. First, participants are given digital recorders and instructed to collect speech from different social occasions. Those people are also provided with different types of microphones. Secondly, some people are given adaptors that can be used to record directly over the telephone. Thirdly, the TV will be a plentiful resource of the Arabic National Corpus. For example, corpus linguists can record the news to study Modern Standard Arabic, to record plays and serials to study different Arabic dialects. Also, we can record sports comments, discussions, commercials, etc.

5.1.2.2. Collecting Written Samples

Collecting samples of writing is the second main task. The first step is to obtain permission from the authors that their writings are included in the corpus. The second step is to gather the written texts, 1,000 to 2,000 word samples that will be computerized.

5.1.3. Computerizing the Corpus

After collecting the written and spoken texts, it can be entered into a database. First of all, we need to transcribe the collected speech. That is representing the oral form of language in a written form. As for English, there are software programs designed to transcribe English speech that has been digitized such as “Voice Walker 2.0”. As for Arabic, we need a program to transcribe Arabic speech and turn it into a spoken form. Concerning computerizing written texts, the first step is to convert written texts into electronic format either with retyping texts or with optical scanners.

5.1.4. Analyzing the Corpus

After computerizing the corpus, the next step is to find the appropriate software programs as well as the appropriate statistical tests for both quantitative and qualitative analysis. These programs and tests are used by linguistic research to analyze the data. As for programs, software programs can be used in corpus analysis.

The most common software program to be used with a corpus is the Concordancing program. Kettemann (1995: 4) argues that the concordancing program is "an extremely powerful hypothesis testing device". This program can be used by researcher to conduct searches for words, group of words, suffixes, prefixes and calculating the frequency. The next step is to subject the information obtained through the programs to some kind of statistical analysis to make frequency counts as well as determining the similarities and differences and to show to what extent they are statistically significant.

Concluding Remarks

Creating the Arabic National Corpus is not an effort of an individual or even a group of individuals, rather it is a national project that needs the collaboration of many institutions in different Arabic countries such as the Arabic language academies, Arab Scientific Research Councils Unions, King Abdul Aziz City For Science and Technology, Arab universities including faculties of arts, faculties of education, and faculties of computers and information systems. Moreover, since it is a national project it needs the governmental support particularly for funding. Creating an Arabic National Corpus will be rewarding and will help in advance the study of the Arabic language and linguistics. Although this work, if taken seriously, will be in its infancy in the Arab world and requires methodological refinement, it seems to be an interesting and promising area of studying Arabic linguistics.

References

Arts, B. (1992) *Small clauses in English: the nonverbal types*. Berlin and New York: Mouton de Gruyter.

Abd-Elghany, A. (1970) *Al-wahadaat al-sarfiyya wa dawraha fi binaa' al-kalima al-Arabiyya* [Morphological units and their roles in Arabic word-formation]. A Published M.A. Dissertation. Cairo: Dar El-Nashr Press.

Abd-Eltawab, R. (1997) *Al-madkhal ila 'lm al-lugah wa manahij al-bahth al-lughawy* [An Introduction to linguistics and methods of linguistic research]. 3rd edn. Cairo: Maktabt Al-Khanji.

Al-Eryaan, M. A. (1984). *Al-maajim al-Arabiyya al-mujanasah* [The hybrid Arabic dictionaries]. Cairo: Dar Al-Muslim.

Al-Trabulsi, M. A. (1996). *Khasa's al-'uslup fi al-shawqiyyat* [Stylistic features of al-shawqiyyat]. Cairo: Al-Majlis Al-Ala Lil-thaqafa.

Aston, G. and L. Burnard (1998). *The BNC handbook: exploring the British national corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, B., T. Sue and B. Levin (1995). Building on a corpus: a linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography* 8.2: 85-114.

Barlow, M. (1999). MonoConc 1.5 and PraConc. *International Journal of Corpus Linguistics* 4.1: 319-27.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8: 241-57.

--- and J. Burgues. (2000). Historical change in the language use of women and men: gender differences in dramatic dialogue. *Journal of English Linguistics* 28.1: 21-37.

Collins, P. (1991). The modals of obligation and necessity in Australian English. In Aijmer and Altenberg. 145-65.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing* 8: 259-65.

Fillmore, C. (1992). Corpus linguistics or computer-aided armchair linguistics. In Svartvik (ed). 35-60.

Fischer, W. (2000). *A grammar of classical Arabic*. 3rd ed. Jonathan Rodgers (Trans.), New London: Yale University Press.

Francis, W. N. (1992). Language corpora B.C. In Svartvik (ed). 17-32.

Gadalla, H. A. et al. (1998). Callhome Egyptian Arabic lexicon. Linguistics Data Consortium, University of Pennsylvania.

--- (1999). *A comparative morphological study of standard Arabic and Cairene Arabic with an analysis of phonological alternations*. PhD dissertation, Assiut University.

--- (2003). *Translating Arabic perfect verbs into English: a text-based approach*. Bulletin of the Faculty of Arts 12: 1-24.

Gavioli, L. (1997). Exploring texts through the concordancer: guiding the learner. In Anne Wichmann, et al. (eds.) 38-99.

Greenbaum, S. (1996). *The Oxford English grammar*. Oxford: Oxford University Press.

Johansson, S. and J. Ebeling. (1996). Exploring the English-Norwegian parallel corpus. In C. Percy, Ch. Meyer, and I. Lancashire (eds.), *Synchronic Corpus Linguistics*. Amsterdam: Rodopi. 3-15.
Kebbe, M. Z. (2000). *A transformational grammar of modern literary Arabic*. London: Kegan Paul International Limited.
Kettemann, B. (1984). On the Use of Concordancing. In ELT. *TELL and CALL* 4: 4-15.
Landau, S. (1984). *Dictionaries: the art and craft of lexicography*. New York: Charles Scribner.
Leech, G. (1992). Corpora and theories of linguistic performance. In Svartvik (ed). 105-22.
--- and M. Short. (1981). *Style in fiction*. London: Longman.
MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk*. 3rd edn. Mahwak, NJ: Erlbaum.
Mahmoud, A. T. (1989). *A comparative study of middle and inchoative alternations in Arabic and English*. Unpublished doctorial dissertation. University of Pittsburgh.
Mansour, M. A. (1999). *The morphosyntactic features of the English and classical Arabic verb: a contrastive study*. MA Thesis, Assiut University.
Meyer, C. F. (2002). *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.
Powell, C. and R. Simpson. (2001). Collaboration between corpus linguistics and digital librarians for the MICASE web search interface. In R. Simpson and J. Swales. 32-47.
Rissanen, M. (2012). Corpora and the study of English historical syntax. In M. Kytö (ed.). *English Corpus Linguistics: crossing paths*. Amsterdam/New York: Rodopi. 197-220.
--- (1992). The diachronic corpus as a window of the history of English. In Svartvik (ed). 185-205.
--- (2000). The world of English historical corpora: from Cædmon to the computer age. *Journal of English Linguistics* 28.1: 7-20.
Schmied, J. and H. Schäffler (1996). Approaching translationese through parallel and translation corpora. In C. Percy, C. Meyer and I. Lancashire (eds), *Synchronic Corpus Linguistics*. Amsterdam: Rodopi. 41-55.
Skaffari, J. (2009). *Studies in early middle English loanwords: norse and French influences* (Anglicana Turkuensia 26). Turku: Department of English, University of Turku.
Stenström, A. and G. Andersen. (1996). More trends in teenage talk: a corpus-based investigation of the discourse item *cos* and *innit*. In C. Percy, C. Meyer and I. Lancashire (eds), *Synchronic Corpus Linguistics*. Amsterdam: Rodopi. 189-203.
Wafi, A. A. (2000). *Fiqh al-lugah* [The philology of the language]. The 6th of October City: Nahdit Misr Lil-Tibaah wa Al-Nashr Wa Al-Tawzee'.

Appendix: The Most Common English Corpora

American Publishing House for the Blind Corpus
Bank of English Corpus
Birmingham Corpus
Brown Corpus
Cambridge Learners' Corpus
Cambridge international Corpus
Canterbury Project
Corpus of Early English Correspondence
Corpus of Middle English Prose and Verse
Corpus of Spoken Professional English
Hong Kong University of Science and Technology Learner Corpus
International Corpus of English (ICE)
International Corpus of Learner of English
Lancaster Corpus
The Northern Ireland Transcribed Corpus of Speech
Talk Bank Project
The Electronic *Beowulf*
Wellington Corpus