Building an International Corpus of Arabic (ICA): Progress of Compilation Stage

Sameh Alansary[*†] Sameh.alansary@bibalex.org
Magdy Nagi[*††] magdy.nagi@bibalex.org
Noha Adly[*††] noha.adly@bibalex.org

[*] Bibliotheca Alexandrina, P.O. Box 138, 21526, El Shatby, Alexandria, Egypt.

[†] Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, El Shatby, Alexandria, Egypt.

[††] Computer and System Engineering Dept. Faculty of Engineering, Alexandria University, Alexandria Egypt.

Abstract

This paper focuses on three axes. The first axis gives a survey of the importance of corpora in language studies e.g. lexicography, grammar, semantics, Natural Language Processing and other areas. The second axis demonstrates how the Arabic language lacks textual resources, such as corpora and tools for corpus analysis and the effected of this lack on the quality of Arabic language applications. There are rarely successful trials in compiling Arabic corpora, therefore, the third axis presents the technical design of the International Corpus of Arabic (ICA), a newly established representative corpus of Arabic that is intended to cover the Arabic language as being used all over the Arab world. The corpus is planned to support various Arabic studies that depends on authentic data, in addition to building Arabic Natural Language Processing Applications.

1 Introduction

There are two points of view about the need for corpora. The first one says that there can not be any corpora, however large, that contain information about all of the areas of any language lexicon and grammar of that language. The second point of view is that every corpus, however small, has taught the person facts that could not be imagined finding out about in any other way.

A number of large corpora have been assembled in the last few years, but the idea itself is not new. It can be traced back to the German linguist Kading, who in 1897 used a large corpus of German - 11 million words - to collate frequency distributions of letters and sequences of letters. This corpus, by size alone, is impressive for its time, and compares favourably with more modern corpora (McEnery and Wilson, 1996). In the late 20th century, there were attempts to develop a number of large corpora of language which can be electronically searched. Early signs of the modern era of corpus linguistics are dated by McEnery and Wilson (1996) from 1960 when Quirk (1960) planned and implemented his ambitious Survey of English Usage (SEU). At the same time, Francis and Kucera began compiling the Brown corpus, which was developed over the following two decades. In 1975 Jan Svartvik began extending the work of the SEU and the Brown corpus to construct the London-Lund corpus. In 1994 the Cambridge and Nottingham Corpus of Discourse in English (Cancode) began, funded by Cambridge University Press and presently contains 5 million words of transcribed data.[1]

In section 2, the roles which corpora can play in a number of different fields of study related to language are discussed. The focus is on the conceptual issues of why

[1] http://www.brown.uk.com/teaching/city/datadriven.pdf

corpus data are important to these areas, and how they can contribute to the advancement of knowledge in each, providing real examples of corpus use. In section 3, the previous trials to building Arabic corpora and the goal of building those corpora are discussed. Also, the advantages and disadvantages of those corpora and tools which are built to analyze those corpora are mentioned. According to previous sections, it is found that there is an urgent need for having a real Arabic corpus, which is discussed in section 4. The focus will be on the International Corpus of Arabic (ICA), its goal and its design and compilation. Section 5 explains the ICA Software program; as ICA would be used in various Arabic language studies, a software application were build to be used for that purpose.

2 Applications of corpora in language research

The importance of corpora to language and linguistics studies is aligned to the importance of empirical data. Empirical data enable the linguist to make objective statements, rather than those which are subjective, or based upon the individual's own internalized cognitive perception of language. As language and linguistics studies cannot rely on intuition or small samples of language; they require empirical analysis of large database of texts as in the corpus-based approach. Corpus-based methods can be used to study a wide variety of topics within linguistics. Because Corpora consist of texts, they enable the linguists to contextualize their analyses of language; corpora are very well suited to more functionally based discussions of language and linguistics. Modern computers have made it possible to store a large number of texts and to analyze a large number of linguistic features in those texts.

However, Linguists of all persuasions have discovered that corpora can be very useful resources for pursuing various resources of research agendas. Due to the importance of corpus in language and linguistic studies, as mentioned in the previous paragraphs, the next five sections will clarify it into five points; the need for corpus in lexicography, grammar, semantics, natural Language Processing and other language studies.

2.1 Corpus in Lexicography

Lexicography research investigated the meanings, synonyms and the use of words. In more recent times, such investigations were extended using corpus-based techniques to study the ways that words are used, considering issues such as:

- How common are different words?
- How common are the different senses for a given word?
- Do words have systematic associations with other words?
- Do words have systematic associations with particular registers or dialects?

By consulting a corpus, the lexicographer can be more confident that the results obtained reflect the actual meaning of a particular word more accurately. Because corpus data contains a rich amount of textual information - regional variety, author, date, genre, part-of-speech tags, etc., it is easier to tie down usages of particular words or phrases as being typical of particular regional varieties, genres and so on. Modern dictionaries depend on corpora of different sizes and types for frequency listings, concordances and collocations, illustrative sentences and grammatical information. If corpora are not representative of the different language usages of a speech community,

they may prove to be unreliable sources of lexicographic information. Corpora have changed the way in which linguists can look at language.

The open-ended (monitor) corpus has its greatest role in dictionary building as it enables lexicographers to keep on top of new words entering the language, or existing words changing their meanings because it is accurately reflecting the actual meaning of a particular word. However, finite corpora also have an important role in lexical studies - in the area of quantification a linguist who has access to a machine readable corpus can call up all the examples of a word or phrase from many millions of words of text in a few seconds. Dictionaries can be produced and revised much more quickly than before, thus providing up-to-date information about language. Also, definitions can be more complete and precise since a larger number of natural examples are examined.

2.1.1 Examples from Arabic show the need for Corpus in lexicography:

A. Investigating word meanings:

One of the advantages of corpus-based research is that the corpus can be used to show all the contexts in which a word occurs. It is possible to identify the different meanings associated with a word, due to one word may have more than one meaning in different contexts, by using corpus this kind of ambiguity can be authentically detected. Table 1 shows the Arabic word "قلب" which has 3 meaning as a noun:

Word meaning	Sentence
core	١. مصر في ***قلب*** الأحداث
heart	٢. أجرى عملية ***قلب*** مفتوح
center, middle	٣. كانت الشمس في***قلب*** السماء

Table 1: The meanings of "قلب" as a noun.

B. Investigating word frequency:

In the first step in understanding the patterns of use associated with a word, some questions should be answered like: which words are the most common in a language? Which are uncommon? Where does a particular word fit on the continuum of very common to very uncommon? In dialectal Arabic telephone speech corpus a list of high-frequency colloquial words has been made.[2] Also, in Lob corpus a list of words frequency has been made (Biber et al.'s 1998).

C. Investigating the variations in lexical category:

In this section, the frequency of a word as a lexical category (noun, verb, adjective, etc.) has been investigated; due to, it may found one word has more than lexical category in different contexts as shown in table 2.

Word meaning	Word Category	Sentence
Ain	Proper-Noun	١. جامعة ***عين*** شمس
wellspring	Noun	٢. ***عين*** الماء
eye	Noun	٣. ***عين*** الإنسان
delimitate/be delimitate	Verb/passive Verb	٤. ***عـــين*** وزيـــرا للخارجية

Table 2: The lexical categories of "عين" word.

[2] http://papers.ldc.upenn.edu/NEMLAR2004/Dialectal-Arabic-telephone-speech-corpus.ppt#304,23,MSA-based orthography

D. Investigating the use of synonyms:

Languages have many words that are considered synonymous. Through corpus, the researchers can easily know synonyms of a word, the frequency of each word of those synonyms and which one of them is more common.

E. The word form according to its case:

As the form of some Arabic words may change according to their case modes (nominative, accusative or genitive), corpus enables the researchers to know the variations which happen to words. For instance the Arabic word "المصريون" changes according to its case as Table 3 shows:

Case	Arabic Word
Nominative	المصريون
Accusative/genitive	المصريين

Table 3: The different cases of "المصري".

2.1.2 Examples of corpus-based lexicography:

In the early 1980, the publisher Collins and Brimingham university compiled the first mega-size corpus, the Cobuild corpus, for a production of a new English dictionary. It produced a number of dictionaries based on two monitor corpora: the Birmingham Corpus (created in 1980 cf. Renouf 1987 and Sinclair 1987, its size at the time 20 million words), and the Bank of English Corpus. By the time the dictionary published in 1987 (Collins 1987).

The Bank of English Corpus has many potential uses, but it was designed primarily to help in the creation of dictionaries. Sections of the corpus were used as the basis of the BBC English Dictionary, a dictionary that was intended to reflect the type of vocabulary used in news broadcasts such as those on the BBC (Sinclair 1992). The vocabulary included in the dictionary was based on sections of the Bank of English Corpus containing transcriptions of broadcasts on the BBC (70 million words) and on National Public Radio in Washington, DC (10 million words). The Bank of English Corpus was also used as basis for a more general purpose dictionary, the Collins COBUILD English Dictionary. Other projects have used similar corpora for other types of dictionaries (Meyer 2002).

The Cambridge Language Survey has developed two corpora, the Cambridge International Corpus and the Cambridge Learners' Corpus, to assist in the writing of a number of dictionaries, including the Cambridge International Dictionary of English. Cambridge Advanced Learner's Dictionary used notes based on the Cambridge Learner Corpus to help students avoid common mistakes (Meyer 2002).

Longman publishers assembled a large corpus of spoken and written American English to serve as the basis of the Longman Dictionary of American English, and used the British National Corpus as the basis of the Longman Dictionary of Contemporary English (Meyer 2002).

The York Junior ELT *(English Dictionary for Learners of English)* dictionary explains the meanings of English words and phrases, and illustrates them with examples drawn from an electronically held corpus of written and spoken texts of modern English. ELT corpus lexicography consists of the compilation of dictionary

entries in a computerized editing system, using an electronic corpus as evidence. In a trilingual dictionary Yilumbu-French-English items are included according to the word tradition and on account of their usage frequency in the corpus[3].
The Oxford Dictionary of English is at the forefront of language research, focusing on English as it is used today, informed by the most up-to-date evidence from the largest language research programme in the world, including the 800-million-word Oxford English Corpus [4].

2.2 Corpus in Grammar

Corpus-based research can be applied to grammar on the word level, sentence level, and discourse level to understand the structure of a language. It is possible to use corpora to obtain information on the structure and usage of many different grammatical constructions and to use this information as the basis for writing a reference grammar of any language. Although grammatical research within linguistics is almost exclusively descriptive rather than prescriptive, it generally does not use empirical methods to study language usage. Descriptive grammarians use field methods to identify the various paradigms in a language, while theoretical grammarians typically rely on their own intuitions about language, sometimes supplemented by asking native speakers to judge whether made-up sentences are grammatically correct or not. Furthermore, none of those approaches focus on variation in language use. Because of this point of view, the grammarians find the important role of corpus in investigating the grammar of a language.

The areas that traditional studies have neglected turn out to be the strengths of corpus-based studies of grammar. The availability of large corpora and computer tools makes it possible to study the patterned ways in which language users use the grammatical resources of a language – by investigating the frequency distribution of various constructions, the association patterns between grammatical structures and other linguistic and non-linguistic factors, and the factors that affect choices between structural variations.

Grammatical (or syntactic) studies, along with lexical studies, were the most frequent types of research which used corpora. Corpora makes a useful tool for syntactical research because of:

- The potential for the representative quantification of a whole language variety.
- Their role as empirical data for the testing of hypotheses derived from grammatical theory.

2.2.1 Examples from Arabic show the need for Corpus in Grammar:

A. Investigating morphological characteristics:

Studying a morphological characteristic in a corpus can teach us both about the frequency and distribution of the characteristic and about the differing functions of particular variations. Morphological analysis by using corpus allows the user to search particular prefixes or suffixes in Arabic language and also infixes may be added. The word "علم" for example may give various meanings by adding different prefixes or suffixes as shown in table 4:

[3] http://www.wat.co.za/Engelse%20Webwerf/Publications/Lexikos/lex16.htm
[4] http://www.askoxford.com/oec/mainpage

Meaning	Suffix	Infix	Prefix	Word
Scientific	ية	***	***	علمية
Learned us	تنا	***	***	علمتنا
His science	ه	***	***	علمه
Scientists	اء	***	***	علماء
Teaching	***	ي	ت	تعليم
Sciences	***	و	***	علوم

Table 4: The various meanings of "علم"word when adding different prefixes or suffixes.

B. The distribution and function of a syntactic construction:

If the grammarians have a large corpus, they can easily determine the distribution of words. For instance, the prepositions "في" , "على" or "من" usually come before and after which word class; verb or noun, after verbs which word class is commonly to occur and so on. By determining such things, the grammarians can rule out the syntax restrictions of that language.

2.2.2 Examples of corpus-based grammars:

There is a long tradition in English studies, dating back to the nineteenth and early twentieth centuries, to use some kind of corpus as the basis for writing a reference grammar of English, a tradition followed by grammarians as Otto (1909–49) or Curme (1947), who based their grammars on written material taken from the work of eminent English writers.

One of the first major reference works to use corpora were the two grammars written by Quirk, Greenbaum, Leech, and Savartvik: *A Grammar of Contemporary English* (1972) and *A Comprehensive Grammar of the English Language* (1985). In many sections of these grammars, discussions of grammatical constructions were informed by analyses of the London Corpus. For instance, Quirk et al.'s (1985) description of the noun phrase concludes with a table presenting frequency information on the distribution of simple and complex noun phrases in various genres of the London Corpus (Meyer 2002).

At Nijmegen University, for instance, primarily rationalist formal grammars are tested on real-life language found in computer corpora (Aarts 1991). The formal grammar is first devised by reference to introspective techniques and to existing accounts of the grammar of the language. The grammar is then loaded into a computer parser and is run over a corpus to test to what extent it accounts for the data in the corpus. The grammar is then modified to take account of those analyses which it missed or got wrong (McEnery and Wilson 2001).

Many smaller-scale studies of grammar using corpora have included quantitative data analysis (for example, Schmied's 1993 study of relative clauses). Now, there is a greater interest in the more systematic study of grammatical frequency, for example, Oostdijk and de Haan (1994) are aiming to analyze the frequency of various English clause types.

Greenbaum's Oxford English Grammar (1996) is based almost entirely on grammatical information extracted from the British Component of the International Corpus of English (ICE-GB), (Meyer 2002). The Collins COBUILD project has created a series of reference grammars for learners of English that contains examples drawn from Bank of English Corpus (Sinclair 1987). Biber et al.'s Longman Grammar of Spoken and Written English (1999) is based on the Longman Spoken and Written English Corpus (approximately 40 million words and contains samples of spoken and written British and American English (Meyer 2002).

Floresta Sintá(c)tica (Portuguese treebank project) is a sampler of c.1,000 running text sentences (European Portuguese). The sampler is a manually revised part of a larger tree corpus (1 million words), which was automatically annotated with the Constraint Grammar based PALAVRAS parser and then converted into constituent trees. This full version can also be searched. The project is a joint venture of the VISL project (Southern Denmark University) and the project "Computational Processing of Portuguese"[5].

2.3 Corpus in Semantics

The studies of semantics of natural languages have yielded a number of interesting results for several semantic phenomena, but it is not clear whether such formal semantic theories can yet be used to automatically process real corpora. Moreover, semantic theories are often based on higher-order formal logics and do not explicitly address the problem of efficient implementations. A new approach that emerged over the past few years is to derive word meaning resources from corpora. Semantic information ranging from synonymy or hyponymy to rather complex verb relations can be learned with a surprising degree of success from corpora. The main contribution that corpus linguistics has made to semantics is by helping to establish an approach to semantics which is objective, and takes account of indeterminacy.

Mindt (1991) demonstrates how a corpus can be used in order to provide objective criteria for assigning meanings to linguistic terms. Mindt points out that frequently in semantics, meanings of terms are described by reference to the linguist's own intuitions. Mindt argues that semantic distinctions are associated in texts with characteristic observable contexts-syntactic, morphological and prosodic. By considering the environments of the linguistic entities, an empirical objective indicator for a particular semantic distinction can be arrived at (McEnery and Wilson 2001).

Another role of corpora in semantics has been in establishing the notions of fuzzy categories more firmly. In theoretical linguistics, categories are usually seen as being hard and fast - either an item belongs to a category or it does not. However, psychological work on categorization suggests that cognitive categories are not usually "hard and fast" but instead have fuzzy boundaries, so it is not so much a question of whether an item belongs to one category or the other, but how often it falls into one category as opposed to the other one. In looking empirically at natural language in corpora, it is clear that this "fuzzy" model accounts better for the data: clear-cut boundaries do not exist; instead there are gradients of membership which are connected with frequency of inclusion.

2.4 Corpus in Natural Language Processing

There are many corpus linguists whose interests are more computational than linguistic. These linguists have created and used corpora to conduct research in an area of computational linguistics known as Natural Language Processing (NLP). For instance, the North American Chapter of the Association for Computational Linguistics regularly has workshops and special sessions at which computational linguistics in NLP discuss the use of corpora to advance research in such areas as tagging, parsing, information retrieval, and the development of speech recognition

[5] http://devoted.to/corpora/

systems. Because researchers in NLP have their own distinct interests, the corpora they use are designed differently than corpora such as Brown or LOB.

The limitations of the corpora created by researchers in NLP, the tool they have developed – taggers and parsers in particular – have been instrumental in creating grammatically analyzed corpora that are of great value to descriptive linguists. For instance, work done by computational linguists in developing the Nijmegen tagger and parser for the International Corpus of English greatly facilitated the grammatical annotation of the British component of ICE, and led to a fully tagged and parsed one-million-word corpus of spoken and written English that, when used with a text retrieval program (ICECUP), can extract from ICE-GB a wealth of grammatical information that in the past could be obtained only by manual analysis.

The USREL team at Lancaster University consist of a group of descriptive and computational linguists who worked together not only to create the British National Corpus but to develop the tagger (CLAWS) that was used to tag the corpus (Meyer 2002).

BulTreeBank project (HPSG-based Syntactic Treebank of Bulgarian) creates a high quality set of syntactic structures of Bulgarian sentences within the framework of HPSG It aims to contain samples of all the syntactic structures of the language. These sentences should serve as templates for future corpora development, become the basis for the development of a more comprehensive test suite for NLP applications They can also be used as a source for grammar extraction and for linguistic research.

2.5 Corpus in other researches

In sociolinguistics, the primary focus is how various sociolinguistic variables, such as age, gender, and social class, affect the way that individuals use language. Corpus-based techniques make it much easier to carry out comprehensive register studies. Although corpora have not yet been used to a great extent in sociolinguistics, there is evidence that this is a growing field. To study variations by gender, in spontaneous dialogue, it becomes necessary to extract from a series of conversations in a corpus what is spoken by males as opposed to females. The analyst might want to consider not just which utterances are spoken by males and females, but whether an individual is speaking to a male or female.

Some registers can be very specific, such as in literature writings, or Methods sections in political research articles. Other registers are more general, such as conversation or student essays. Difficulties arise without corpus-based techniques because comprehensive register studies have three important requirements:

- inclusion of a large number of texts; because register studies based on too few texts are likely to be inaccurate,
- consideration of a wide range of linguistic characteristics; because register studies based on a few range of linguistic characteristics do not provide comprehensive register descriptions, and generalization based on such studies are likely to be inaccurate,
- and comparison across registers; because a baseline for comparison was needed to know whether the use of a linguistic feature in a register is rare or common.

Kjellmer (1986), for example, used the Brown and LOB corpora to examine the masculine bias in American and British English. He looked at the occurrence of masculine and feminine pronouns, and at the occurrence of the items man/men and woman/women. Another hypothesis of Kjellmer's was not supported in the corpora - that woman would be less "active", that is would be more frequently the objects rather than the subjects of verbs. In fact men and women had similar subject/object ratios.

Holmes (1994) makes two important points about the methodology of these kinds of study, which are worth bearing in mind. First, when classifying and counting occurrences, the context of the lexical item should be considered. Second, Holmes points out the difficulty of classifying a form when it is actively undergoing semantic change (McEnery and Wilson 2001).

In BNC, to enable the study of sociolinguistic variables in the spoken part of it, each conversation contains a file header and a statement at the start of the sample providing such information as the age and gender of each speaker in a conversation. A software program, Sara, is designed to read the headers and do various analyses of the corpus based on a pre-specified selection of sociolinguistic variables.

In the British component of the International Corpus of English (ICE-GB), ethnographic information on speakers and writers is stored in a database, and a text analysis program is designed to analyze the corpus, ICECUP (Meyer 2002).

Also, geographical variations are considered as corpora have long been recognized as a valuable source of comparison between language varieties as well as for the description of those varieties themselves. Certain corpora have tried to follow as far as possible the same sampling procedures as other corpora in order to maximize the degree of comparability. For examples, the LOB corpus contains roughly the same genres and sample sizes as the Brown corpus and is sampled from the same year (i.e. 1961). The Kolhapur Indian corpus is broadly parallel to Brown and LOB, although the sampling year is 1978 (McEnery and Wilson 2001). One of the earliest pieces of work using the LOB and Brown corpora was the production of a word frequency comparison of American and British written English. These corpora have also been used as the basis of more complex aspects of language such as the use of the subjunctive (Johansson and Norheim 1988).

Few examples of dialect corpora exist at present - two of which are the Helsinki corpus of English dialects and Kirk's Northern Ireland Transcribed Corpus of Speech (NITCS). Both corpora consist of conversations with a fieldworker - in Kirk's corpus from Northern Ireland, and in the Helsinki corpus from several English regions. Such elicitation experiments tend to focus on vocabulary and pronunciation, neglecting other aspects of linguistics such as syntax. Dialect corpora allow these other aspects to be studied (McEnery and Wilson 2001).

Also, corpus examples are important in language learning as they expose students to the kinds of sentences that they will encounter when using the language in real life situations. Resources and practices in the teaching of languages and linguistics tend to reflect the division between the empirical and rationalist approaches, the non-empirically based teaching materials can be misleading and that corpus studies should be used to inform the production of material so that the more common choices of usage are given more attention than those which are less common. Research in language acquisition and development has focused on:

1. the first-language acquisition of very young children,
2. later language development, such as the acquisition of literacy skills, by students at various stages,
3. and second language acquisition, by children and adults.

There are three advantages of the corpus-based approach in language acquisition and development:

1. By including a relatively large collection of texts, corpus-based analyses provide the basis for generalizations concerning groups of speakers and writers at different stages.

2. By investigating the association patterns among sets of linguistic features, corpus-based analyses enable more comprehensive descriptions of language use at different developmental stages.
3. By including texts from multiple registers, corpus-based analyses facilitate broader perspectives on language development.

Now there are corpora suitable for studying both first- and second-language acquisition. Kennedy (1987a, 1987b) has looked at ways of expressing quantification and frequency in ESL (English as a second language) textbooks. Holmes (1988) has examined ways of expressing doubt and certainty in ESL textbooks, while Mindt (1992) has looked at future time expressions in German textbooks of English. These studies have similar methodologies - they analyze the relevant constructions or vocabularies, both in the sample text books and in standard English corpora and then they compare their findings of the two sets (McEnery and Wilson 2001).

Corpora have also been used in the teaching of linguistics. Kirk (1994) requires his students to base their projects on corpus data which they must analyze in the light of a model such as Brown and Levinson's politeness theory or Grice's co-operative principle. In this approach, Kirk is using corpora not only as a way of teaching students about variation in English but also to introduce them to the main features of a corpus-based approach to linguistic analysis (McEnery and Wilson 2001).

A further application of corpora in this field is their role in computer-assisted language learning. Recent work at Lancaster University has looked at the role of corpus-based computer software for teaching undergraduates the rudiments of grammatical analysis (McEnery and Wilson 1993). McEnery, Baker and Wilson (1995) carried out an experiment over the course of a term to determine how effective Cytor was at teaching part-of-speech learning by comparing two groups of students - one was taught with Cytor, and the other was taught via traditional lecturer-based methods. In general the computer-taught students performed better than the human-taught students throughout the term (McEnery and Wilson 2001).

To study second-language acquisition, a number of researchers began developing what are called learner corpora: corpora containing the speech or writing of individuals learning English as a second or foreign language. One of the larger corpora in this area is called the International Corpus of Learner English (ICLE). ICLE is currently more than two million words in length. Other learner corpora include the Longman's Learner Corpus and the Hong Kong University of Science and technology (HKUST) Learner Corpus (Meyer 2002).

Due to all the previous discussion about the need for corpus, and as there is no Arabic corpus made to be used in linguistic/language studies as will be shown in the next section, all researchers are asking for a representative and balanced Arabic corpus.

3 Arabic corpora

While it is of prime importance to descriptive corpus linguistics to create valid and representative corpora, in the field of natural language processing (NLP) this is an issue of less concern. Obviously, the two fields have different interests: it doesn't require a balanced and representative corpus to train a parser or speech-recognition system. But it would greatly benefit the field of corpus linguistics if descriptive corpus linguists and more computationally oriented linguists and engineers work together to create corpora. In the past several years, on-line corpora became increasingly accessible, and corpus-based studies became very common. The British National

Corpus is a good example of the kind of corpus that can be created when linguists, computational linguists, and publishing industry cooperate.(Meyer(2002))

Natural language processing (NLP), including Information Retrieval, Machine Translation and other Natural Language-related disciplines, is showing more interest in the Arabic language in recent years. Suitable resources for Arabic are becoming a vital necessity for the progress of this research. Corpora are an important resource but Arabic lacks sufficient resources in this field, so a research projects need to compile a corpus, which represents the state of the Arabic language at the present time and the needs of end-users. Therefore many trials have been conducted to build Arabic corpora but some of them were unsuccessful trials and others were for commercial purposes.

3.1 Some trials in compiling Arabic corpora

In this section, we will present some previous trials of compiling Arabic corpora that were collected for several purposes of applications for Arabic.

3.1.1 CLARA (Corpus Linguae Arabicae)

CLARA corpus is an electronic corpus of Modern Standard Arabic; it contains 37 million words. Although it surveys Modern Standard Arabic, some 300 000 words come from Western languages, mostly English. More than 95% of texts come from last ten years. CLARA corpus is collected to reflect the contemporary language usage but with two exceptions of text sources: Qur´an and Bible. Geographically, the areas covered are the Arabic Peninsula, Syria and Egypt, with some examples from other countries, like Tunisia and Morroco, with low percentages which do not exceed 5%. From the content point of view, the share of popular fiction texts is about 15%; the rest comes from informative domain, i.e. the corpus aims at terminological richness. As for the medium, about 50% of the texts come from periodicals, 35% from books and 15% from miscellaneous written materials .The final goal is building a balanced and annotated corpus – the annotation should be done for morphological boundaries and POS. The acquisition of texts for the corpus included scanning of Arabic books (especially fiction, but also texts from informative domains, like medicine, astronomy, law, technology, etc,(using the Arabic OCR program Automatic Reader (by Sakhr).

3.1.2 Arabic Newswire Corpus

Newswire Corpus was released in 2001 by David Graff and Kevin Walker. The source material was tagged using TIPSTER-style SGML and was transcoded to Unicode (UTF-8). The corpus includes articles from May 13, 1994 to December 20, 2000.

If we cogitate this corpus, we will find that although it contains 80 M written words, it has no representativeness, diversity or balance for all Arabic language because it was collected from press only such as (Agence France Presse, Xinhua News Agency, and Umma Press). The data is in 2,337 compressed (zipped) Arabic text data files. The corpus is constructed to several applications such as Education and the development of technology.

3.1.3 A corpus of Contemporary Arabic (Poste) (CCA Corpus)

The ACC Corpus was released from University of Leeds (UK) by Latifa Al-Sulaiti & Eric Atwell. Their survey confirms that existing corpora are too narrowly limited in source-type and genre, and that there is a need for a freely-accessible corpus of contemporary Arabic covering a broad range of text-types. Their survey also shows

support for the inclusion of parallel English-Arabic samples. Supplementary questions showed support for potential use in wide range of Language Engineering applications; and indicates that teachers of Arabic as a foreign language already make significant use of computers in teaching, and want to include contemporary, authentic examples.

The researchers invested a good amount of time and effort to annotate every text with a header which provides internal and external information. The information that they included in the header follows largely accepted standard but they included only the minimal required information such as authorship, publication and details about texts such as their types and domains. The minimal components were: File description, Encoding description, and Profile description. In addition, they annotated the texts with paragraphing. The method they used for processing the files was that they created a template of the header and handled them in the Unicode editor UNIRED. Collecting the texts for the corpus and annotating them were all done manually.

What noted above reveals a good effort in this corpus but it has some disadvantages. First, the compiled corpus was only one-million words covering some of the categories they decided to collect. Secondly, it found that this written corpus doesn't contain written materials only but also spoken materials such as: short stories, **radio**, newspapers, children's stories, health and medicine, autobiography, magazines and economics. The third disadvantage is that it does not have an appropriate document-classification; documents were collected without any consideration of representativeness.

3.1.4 Al-Hayat Corpus

The corpus was developed in the course of a research project at the University of Essex, but it's important to make some points clear: the corpus is so small; it is only 18.5 million words. As for representativeness, this corpus does not represent a long period of coverage as the dataset is an electronic archive for the newspaper of the year 1998 only and doesn't cover any other year. According to the articles classification, it covered 7 subject categories based on Al-Hayat classification: General, News, Economic, Sports, Computers and Internet, Science and Technology, Cars and Business.

3.1.5 An-Nahar Newspaper Text Corpus

It is a newspaper text corpus according to An-Nahar Lebanon newspaper that comprises articles in standard Arabic from 1995 to 2000 (6 years) stored as HTML files on CD Rom media. Each year contains 45 000 articles and 24 million words. Each article includes information such as title, newspaper's name, date, country, type, page, etc.[6]

3.1.6 Classical Arabic Corpus (CAC)

This corpus was compiled by Abdel-Hamid Elewa at the University of Manchester Institute of Science and Technology. If we make quick description of this corpus we will find out that this corpus contains a very few number of words (5M words) to survey the major of contemporary Arabic language. It comprises texts including short poems from the period of the advent of Islam up to the end of the eleventh century. The material is derived from the web The main division of the corpus is intended to be between fiction and non-fiction. However, since fiction represents only 11% which is

[6] http://www.elda.org/catalogue/en/text/W0027.html (Copyright © 1996-2001 ELRA/ELDA)

due to unavailability of fictional material for this period, the text types are divided into four genres; thought and belief, literature, linguistics, and science. This corpus is not tagged at this stage as it was mainly developed for the purpose of lexical analysis[7].

3.1.7 General Scientific Arabic Corpus (GSAC)

This corpus was developed by Amin Al-Muhanna at the University of Manchester, Institute of Science and Technology. Its purpose is to investigate how scientific and technical terms are formulated in Arabic with a focus on compounds. In addition, his research compares between the mechanism used by Arab writers and what has been proposed by language academies. This corpus is limited to one country only and doesn't cover the diversity and representativeness of major of Arabic language, the material is derived from the Kuwaiti magazine site 'Science and Technology'. Part of this corpus (1M) has been tagged. Al-Muhanna reported that his training corpus contained 100,000 tokens and the accuracy of his tagging was 92%.

3.1.8 Arabic Gigaword Corpus

Arabic Gigaword corpus released in Jul 22, 2003 by David Graff; Arabic Gigaword is a comprehensive archive of newswire text data that has been acquired from Arabic news sources by the LDC. Each data file name consists of the seven-letter prefix, an underscore character ("_"), and a six-digit date (representing the year and month during which the file contents were generated by the respective news source), followed by a ".gz" file extension, indicating that the file contents have been compressed using the GNU "gzip" compression utility (RFC 1952). Therefore, each file contains all the usable data received by LDC for the given month from the given news source.

4 The need for Arabic tools

Over the past decade, there has been some important progress in the computational processing of Arabic. However, Arabic is still lacking tools and annotated resources. Many researchers in the field attest that fully automated fundamental Arabic NLP tools such as Base Phrase Chunkers are still not available for Arabic (Maamouri M. et al.'s, 2004). Arabic NLP is still in its infancy, due to the problem of obtaining large amounts of text data (Duh K. et al's, 2005). So, native speakers of languages that are not well served by language technology suffer from less access to information, and from less efficient tools, and higher productions costs for documents and translation (Maegaard B. et al.'s, 2005). This section will survey some of the tools needed for processing Arabic corpora, so progress can be achieved in the related area of application.

1. Machine translation:

The idea of using parallel corpora is not new; it dates back to the early days of machine translation, but it was not used in practice until 1984 (Guidère M., 2002). Unfortunately, Arabic lacks such technological development, along with the huge volume of translations available in different languages, which points toward the use of Arabic corpus for specific machine translation and computer-assisted translation applications to permit fine future tools for Arabic corpora.

[7] http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm

2. Lexical Techniques:

Lexical choice (lexicalization) is a process of mapping meaning representations onto lexical items (words) in the language because it is at the heart of generation and having good lexicalization systems is important for systems that will convey ideas in natural languages (e.g., Machine Translation (MT) and Natural Language Generation (NLG) systems)(Al-jabri 1997). So, a generation system in Arabic will need to identify all possible words and choose among them the best candidate in a particular situation by using Arabic corpora. Performing lexical choice is non-trivial because the meaning representations are not directly linked to words and choosing the right word requires knowledge not only about the semantics, but also about syntax and pragmatics.

3. Word segmentation:

By its ideal definition, lemmatization is a process wherein the inflectional and variant forms of a word are reduced to their lemma (Siemens 1996). Due to the complex meanings decomposable into several morphemes (i.e. prefix, stem, suffix) in morphologically rich languages like Arabic, such languages present significant challenges to many NLP applications such as lemmatization which is important to be used in dictionaries and concordances. Lemmatizing a text helps generating word indexes, concordances, and dictionaries from that text with ensuring that all forms of a particular word within it can be existed by searching only for its lemma form.

4. Building POS tagger:

POS tagging of Arabic texts ambiguity needs to be resolved by POS tagger through the use of a statistical language model developed from Arabic corpus, but unfortunately, Arabic is still missing real corpora to apply such useful tool for Arabic.

5. Vocalizing Arabic Texts:

Arabic language has two kinds of vowels: Long vowels which are written as normal letters; and short vowels which are written as punctuation marks, above or below letters. Search engines, text to speech engines, and text mining tools are just some examples of applications that need Arabic texts to be vocalized before being processed (Saady et al.'s, 2006).

6. Building a Stemmer:

Arabic language is in need for building a stemmer in NLP to be used for effective information retrieval. Although, Arabic language presents agglutination of articles (letters), prepositions, and conjunctions at word initial position as well as at the word final position, it lacks a good stemmer.

7. Word Sense Disambiguation for Information Retrieval:

Although recorded information usually is retrieved by means of stored data that represents documents, it is the emphasis on information relevant to a request, rather than direct specification of document, that characterizes the modern subject of information retrieval (Heaps 1978). Consequently, Information Retrieval (IR) systems as data retrieval tools are very important. However, in Arabic language the word sense disambiguation is one of great reasons for poor performance of these systems and their lack of access to the satisfactory level.

9. Text Data Mining:

Text data mining, as a multidisciplinary field involving information retrieval, text analysis, information extraction, clustering, categorization and linguistics, is becoming of more significance, and efforts have been multiplied in studies to provide for

fetching the increasingly available information efficiently (Eldos 2002). Due to the Arabic language lacking of corpora, it is difficult to display textual content and quantitative data of Arabic.

5 The International Corpus of Arabic (ICA)

Corpus-based approaches to language have introduced new dimensions to linguistic description and various applications by permitting some degree of automatic analysis of text. The identification, counting and sorting of words, collocations and grammatical structures which occur in a corpus can be carried out quickly and accurately by computer, thus greatly reducing some of the human drudgery sometimes associated with linguistic description and vastly expanding the empirical basis.

Linguistic research has become heavily reliant on text corpora over the past ten years. Due to the increasing need of an Arabic corpus to represent the Arabic language and because of the trials to build an Arabic corpus in the last few years were not enough to consider that the Arabic language has a real, representative and reliable corpus, it was necessary to build such an Arabic corpus to support various linguistic research on Arabic.

Bibliotheca Alexandrina (BA) is one of the international Egyptian organizations that plays a noticeable role in disseminating culture and knowledge, and in supporting scientific research. It initiated a big project to build the "International Corpus of Arabic (ICA)"; a real trial to build a representative Arabic corpus as being used all over the Arab world to support research on Arabic.

5.1 The goal of ICA

The International Corpus of Arabic (ICA) is planned to contain 100 million words. The collection of samples is of written Modern Standard Arabic selected from a wide range of sources which is designed to represent a wide cross-section of regional variety of Arabic; it is stimulating the first systematic investigation of the national variety as being used all over the Arab world.

ICA is a step-by-step guide to creating and analyzing Arabic linguistic corpora. Demonstrating those corpora have proven to be very useful resources for linguists who believe that their theories and descriptions of Arabic should be based on real, rather than contrived, data.

5.2 Planning the construction of ICA

A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research. It is important to realize that representing a language – or even part of a language – is a problematic task. We do not know the full extent of variation in languages or all the contextual variables that need to covered in order to capture all variation in texts. However, attention to certain issues will ensure that a corpus is as representative as possible, given our current knowledge of language.

In planning the collection of texts for the ICA, a number of decision related to corpus design such as representativeness, diversity, balance and size were taken into consideration:

1. Some sources of texts were determined.
2. A variety of different genres of writing would be gathered for inclusion in the corpus.
3. Each genre would be divided into text samples.
4. A careful record of a variety of variables would be kept, including when and where the texts were written or published; whether they were books, articles, or net articles.
5. Decisions related to what the proportions (weighting) between different sections and genres in a corpus have been taken. Such decisions were taken upon how common the genre or the source is (Balance in a corpus is not addressed by having equal amounts of text from different sources and genres).
6. Decisions about how many categories the corpus should contain, how many samples the corpus should contain in each category and how many words there should be in each sample were taken into consideration. Some discussions of size in corpus design focus exclusively on the number of words in the corpus, However, issues of size also relate to the number of texts from different categories, the number of samples from each text, and the number of words in each sample. Enough texts must be included in each category to encompass variation across authors.

Before the texts to be included in a corpus are collected, annotated, and analyzed, it is important to plan the construction of the corpus carefully: what size it will be, what types of texts will be included in it, and what population will be sampled to supply the texts that will comprise the corpus. Ultimately, decisions concerning the composition of a corpus will be determined by the planned uses of corpus. To explore the process of planning a corpus, the following sections will consider the methodological assumptions that guided the compilation of International Corpus of Arabic.

5.2.1 Corpus Design and Compilation

It is important to realize that the creation of ICA is a "cyclical" process, requiring constant re-evaluation as the corpus is being compiled. Consequently, we are willing to change our initial corpus design if circumstances arise requiring such changes to be made. In corpus design and compilation a lot of decisions were taken into consideration to reach to the basic goal of ICA.

5.2.2 Sources and genres included in the ICA

The ICA contains a diverse range of sources. Some of these sources are divided into other sub sources (Table 5):

Sources	Sub sources
Press (A)	**Newspapers**
	Magazine: **General** **Specialized**
	Electronic press
Net Article	
Books	
Academic sources	

Table 5: The sources and sub sources of ICA.

In press, texts were complied from principal daily newspapers in different Arab countries (Table 6):

Country	Newspapers
Egypt	**Ahram**
	Akhbar
	Gomhoria
	Akhbar-elryadah
	Elssyasa Al-Dawlia
Saudi Arabia	**Al-Watan**
	Al-Jazyra
	Hedaya
Qatar	**Al-Rayah**
Jordan	**Addustoor**
Iraq	**Kol El-Iraq**
	Al-Etthad
Palestine	**Al a'maan**
	Al-Quds
	Alhayat.Algadidah
U.A.E "Dubai"	**CNN**
Arab magazine published outside the Arab world	**Al-hayat (London)**
	Sawt-Alkeraza

Table 6: The principal daily press in the Arab world

The main focus of compiling texts was to cover the same genres from different sources in all Arab countries. So it must be noted that the Arab countries in table 2 are not the only countries to be included in the ICA, rather they are the countries included so far, the rest of the Arab countries will be added.

The ICA also contains a diverse range of written genres. Some of these genres are divided into other subgenres, as shown in table 7:

Genres	Sub-genres
Strategic Sciences.	Politics.
	Law.
Social Sciences.	Economy.
	Sociology.
Sports.	
Religion.	Islam.
	Christianity.
	Other religions
	Comparative religion.
Literature.	**Prose:** Novels. Short Stories. Child Stories. Plays.
	Poetry.
	Studies of Literature and Linguistic.
Humanities.	History.
	Psychology.
	Philosophy.
	Geography.
Natural Sciences.	Biology.

	Physics.
	Chemistry.
	Geology and Environment.
Applied Sciences.	Medicine.
	Engineering.
	Agriculture.
	Technology.
Arts.	
Biography.	

Table 7: The genres and sub-genres of ICA

Tables (4&6) indicate the final classification of sources and genres that the compilers of ICA reached so far according to the actual availability of texts.

5.2.3 Design criteria

In designing the ICA some sources and genres were determined according to the following design criteria:

- The design of corpus identified the common sources and genres of the written texts.
- The design depended primarily on sources of texts and in each source there were some genres. This design will make the search in texts more economic and easier.
- Collecting the texts to convey all regional variations of the Arabic Language was the basic interest in the design, for example, newspapers were planned to convey a wide range of sources from different countries; these newspapers tried to exemplify the principal daily newspapers.

Table 8 shows the first version classified according to the texts that will be compiled:

Sources of texts	Type of texts (genres)
Newspapers	**Politics** **Art** **Literature** **Culture** **Sciences** **History** **Economy** **Religion** **Sport** **Sociology** **Miscellanies**
Magazine	**Politics** **Art** **Literature** **Culture** **Science** **History** **Economy** **Religion** **Sport** **Sociology**

	Miscellanies
Novels	**Fictions** **Un Fictions** **Child Story** **Religion** **Plays**
Books	**Art** **History** **Sport** **Religion** **Sciences** **Culture** **Literature**
Net Articles	**Politics** **Art** **Literature** **Sciences** **History** **Economy** **Religion** **Sport** **Miscellanies**
Academic	**Sciences**

Table 8: The first classification of ICA.

Once the texts representing genres were actually began to be collected, some problems were contradicted with the classification in table 4 were faced:

- It is not easy to find texts in a specific genre; not every genre that exists in theory exists in practice. For example, in newspaper and magazine it is found that the historical, cultural and scientific texts were very few, therefore, the history in newspaper should not be considered as a separate genre under the newspaper source, otherwise, the number of texts representing each genre will not be balanced. Such a few number of texts in each genre under the newspaper source were placed under miscellanies.
- Some texts could not be categorized easily i.e. either the text genre is not clear or the same text may be related to more than one possible genre. For example, in books it is found that one book may follow more than one genre; another may talk about politics and economy.
- Some other sources were found out to be missing from the theoretical classification in table 4, e.g. “recipes” and “manual”. Once recipes and manuals have been compiled it is found that these two sources should not be considered as separate sources but as two genres; manuals considered as kind of the source “books”.
- Some other genres are missing from the theoretical classification; for example, in books a “politics” genre was missed .
- Not every source over the Internet is suitable for data collection as many of them contain many spelling mistakes; others do not have reference information, even sometimes the author is not mentioned.
- Many different newspapers may have the same article. The newspapers of **Al-Dustoor** and **Al-Ahram** are daily newspapers published in different countries;

however, it is found that some articles are the same in both of them. To avoid such problems, data were not collected from the newspapers at the same day.

- Many issues of the same newspaper contain the same article; websites of some daily newspapers were not updated everyday which decreases the sampling rate. For example, in **Al-Rayah** and **Al-Hayat** newspapers a few new articles may be added and most of the old articles still exist with the same old date. It is found that in some days there were no new articles to be taken. In **Al-Ahram** newspaper it is found that some articles do not change but the date of those articles are changed daily which causes accidental repetition of texts.
- Websites of some magazines are not updated periodically which obstacles the sampling rate. **Akher-Saʿa** magazine, for instance, was not updated for about 5 months.
- Novels source were considered as books representing a separate genre.
- The availability of free books over the Internet was difficult.

According to the previous design criteria, some modifications were made to the classification in table 4 to represent Arabic according to its authentic use and distribution:

1. Magazines source is sub-classified into general magazine, where articles represent more than one genre, and specialized magazines, where articles discuss only one genre.
2. Newspapers and magazines were included under one source, "Press" under which they are considered sub-sources. Another sub-source that resembles "newspapers" and "magazines" is added, namely "electronic press"; the non-printed press that exists on the Internet only.
3. Some genres were included under larger genres, for example history, psychology, philosophy and geography genres would be included under humanities under which they are considered subgenres.
4. Some genres and sub-genres have been added, e.g. Poetry, Biography, linguistic and literary studies, etc.

The current document hierarchy of the ICA, so far, is :

- There are 4 sources all over the corpus, namely, Press, Net articles, Books and Academics.
- Some sources are divided into sub-sources, namely, Newspapers, Magazine, and Electronic Press (in Press source)
- There are 11 genres all over the corpus, namely, , Strategic Sciences, Social Sciences, Sports, Religion, Literature, Humanities, Natural Sciences, Applied Sciences, Art, Biography and Miscellanies.
- .Some genres are divided into sub-genres, namely, Politics, Law, Economy, Sociology, Islamic, Christian, Other religions, Comparative religion, Novels, Short Stories, Child Stories, Plays, Poetry, Studies of Literature and Linguistic, History, Psychology, Philosophy, Geography, Biology, Physics, Chemistry, Geology and Environment, Space, Medicine, Engineering, Agriculture and Technology.

Figure 1 shows one of main sources (Press) in ICA and one of its sub-sources (Newspapers). In addition, it shows some of genres and sub-genres that were found in the ICA classification:

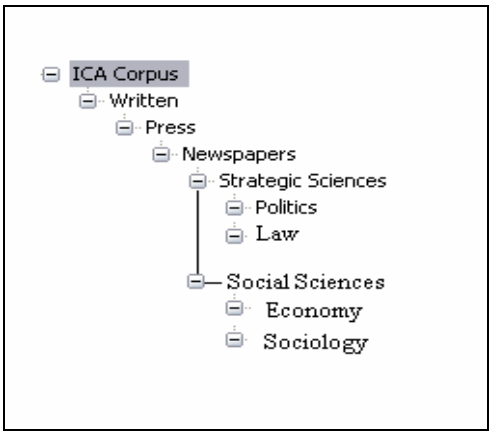

Figure 1: One of main sources in ICA

Table 9 shows the percentage of words that have been collected until now. At this point, we must make a certain issue clear. These percentages are not final as the collection of texts is still under development.

The sources and sub-sources included in ICA written corpus		
Sources	**Sub-sources**	**% of written words**
Press		23.24
	Newspapers	15.64
	Magazine	5.50
	Electronic press	2.10
Net Articles		7.53
Books		69.23
The genres included in ICA written corpus		
Genres	**% of written words**	
STR (Strategic Sciences)		18.03
SOL (Social Sciences)		3.58
REL (Religion)		25.39
LIT (Religion)		37.08
APP (Applied Sciences)		2.26
HUM (Humanities)		2.88
NTU (Natural Sciences)		0.95
ART(Art)		1.00
SPT (Sports)		2.25
BIG (Biography)		0.80
IS (Miscellanies)		5.78

Table 9: The percentages of sources, sub-sources and Genres

5.3 Pre-processing stage

There are a number of general considerations to bear in mind when beginning the process of computerizing the written corpus. The pre-processing stage is an important stage for the compiled texts to be prepared to help in analysis and search processes on the corpus. The collected text needs some pre-processing stages to be prepared.

1. **HTML to TXT conversion.**

   Because the texts to be included in a corpus will be edited with a word-processing program, it may be tempting to save computerized texts in a file format used by a word-processing program (such as files with the

extension.DOC in Microsoft Word). However, it is from the onset to save texts in TXT format, since this is the standard format used for texts included in corpora, and to use a simple text editor to work with texts rather than a standard word-processing program.

2. **Striping HTML codes.**
   Some hidden characters that were found when HTML to TXT conversion happened need to be deleted because such hidden characters may cause problems in both search and analysis stages. So striping HTML codes is an important stage that helps in cleaning up any of these hidden characters from the TXT files.

3. **Coding file names (Meta-information)**
   When creating a corpus, it is practical to save individual texts in separate files stored in directories that reflect the hierarchical structure of the corpus. Organizing a corpus into a series of directories and subdirectories makes working with the corpus much easier, and allows the corpus compiler to keep track of the progress being made on corpus as it is being created. ICA consists of a main directory that containing all the written texts. This directory, in turn, is divided into a series of subdirectories containing the main types of writing that are collected. Each text receives an additional file extension, coding files name with meta information in order to indicate the directory and subdirectory that the text file belongs to.

   It is very important to rename the compiled files according to some parameters to facilitate search processes and to prevent overlapping among file names in case they are placed in the same folder. This coding process of text file names will enable us in the future to customize the search in any part of the corpus. For example, the abbreviations in the filename: AH10-A1.1.1-140207 can be indicated in Table10:

AH10	Contains two pieces of information: Ahram newspaper source, the attached number indicates that this file is the 10^{th} article in that newspaper with the same genre, subgenre and date.
A1.1.1	Contains three pieces of information: Newspaper source (A1), Strategic science "genre" (A1.1) and Politics "sub-genre" (A1.1.1).
140207	Contains three pieces of issuing information: The day (14), the month (02) and the year (2007)

Table 10: An example of codes of filenames

4. **Editing Stage:**
   The editing stage deals with fixing spelling mistakes and grammar mistakes to prepare texts for analysis. The texts with spelling errors cause both of search and analysis problems**.** Most popular problems in spelling mistakes in Arabic are associated with the “Hamzah” and “Yaa'”.

6.1 Markup codes

Metadata is usually defined as 'data about data'. For a corpus to be fully useful to potential users, it needs to be annotated. There are three types of annotation, or

"markup," that can be inserted in a corpus: "structural', "part-of-speech", and "grammatical". [8]

Here the focus will be on the first type of markup. Structural markup provides a descriptive information about the texts. For instance, general information about a text can be included in a "file header", which is placed at the start of text, and can contain information as complete bibliographic citation for a written text, or orthographic information about the participants (e.g. their age and gender) in spoken dialogue. Within the actual spoken and written texts themselves, additional structural markup can be included to indicate, for instance, paragraph boundaries in written texts.

In the ICA there are some structural markup within each text as Table 11 shows:

D/ T/ نقل ملكية عمر أفندي إلى أنوال السعودية /T P/ أخيرا.. تم نقل ملكية شركة عمر أفندي أمس إلى شركة أنوال السعودية . وقرر الدكتور محمود محيي الدين وزير الاستثمار تعيين محمد وهب الله رئيس النقابة العامة ممثلا عن حصة المال العام وقدرها ١٠ % . /P /D

Table 11: Some of structural mark up in one of texts.

/D	The beginning of document.
D/	The end of document.
/T	The beginning of title.
T/	The end of title.
/P	The beginning of paragraph.
P/	The end of paragraph.
/Q	The beginning of question.
Q/	The end of question.

It must be noted that the structural markup is not identical in all corpora systems and the symbols to be included in each corpus are determined by the compilers of corpus

6 ICA Software

A software is built to help researchers to interrogate the corpus. The ICA software is an application that is considered an adequate tool to help in exploring the Arabic language texts. This section provides a detailed description of the available features of software. The ICA software provides an overview of the corpus (shows the document hierarchy of the corpus) and presents all suitable information at the request of researchers interested in authentic data about Arabic. At its current stage, the ICA software has the following capabilities:

1. **Insertion of documents:** It is possible to insert a new document; a single document or a batch of documents (Figure2). The system is able to detect the genre and the source of the document according to its filename code and inserts it under the appropriate node in the document hierarchy.

[8] There is also annotation, such as semantic annotation, that can be used to mark up higher level structures larger than the word, clause, or sentence.

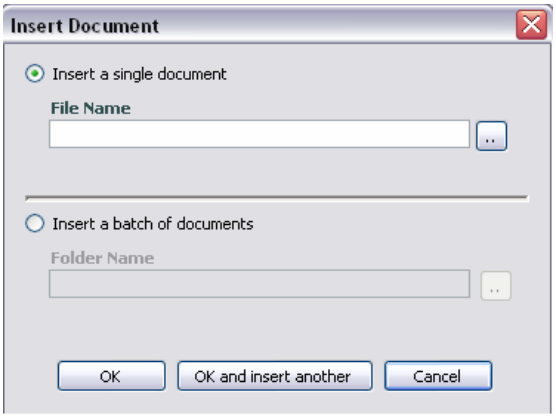

Figure 2: Insertion Window

2. **Searching the corpus:** Different search types and search options are available . The search options enable the user to select the domain of the search in corpus where the user may search in current document and current location marked in the tree of document hierarchy. In addition, the search options enable the user to select the display option to view the results; either in context or separately (Figure3):

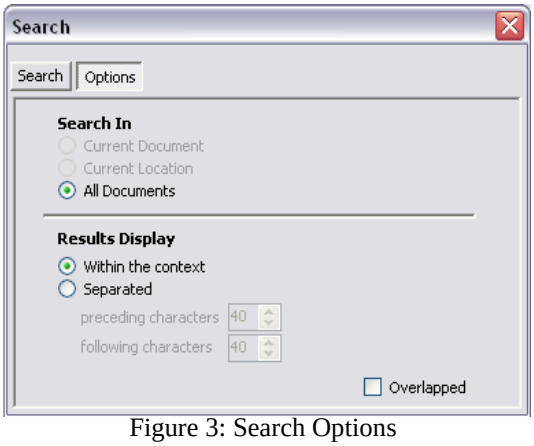

Figure 3: Search Options

Three different types of search are provided: Exact match, Wildcard and Regular Expression (Figure 4).

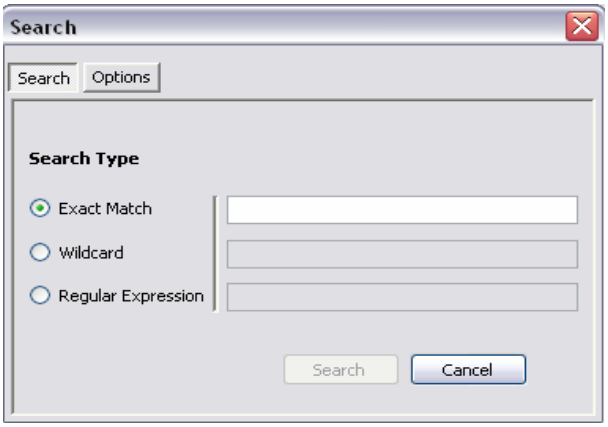

Figure 4: Search types

The difference between these three types of search is indicated in the following example, if two different characters are selected, for example "فـي", these two

characters would be used in different ways in the three types of search and the results will differ from one type to another:

- In wildcard search type the user can write these two characters, in sequent or separated by (* or ?). The (*) means any sequent of characters after or between these two characters, therefore if the user searches for (ف*ي) the search results will include any word containing these two characters and any string of character between them. The (?) means zero or one character after or between these two characters, therefore, if the user searches for (في??) the search results will include any word that contains the sequence of these two characters and followed by any two characters i.e. any word of 4 characters in length beginning with (في).
- Regular Expression search is the most powerful type of search; it enables the user to search for any type of strings even above word level. For example, "{2}[ف-ي]" means searching about any two characters from (ف) to (ي).

In separated context options the user must decide the number of preceding and following characters before and after the keyword. A sample of output of separated search can be seen in figure5. It is also possible for the output to be presented in a whole text browsing view as seen in figure 5.

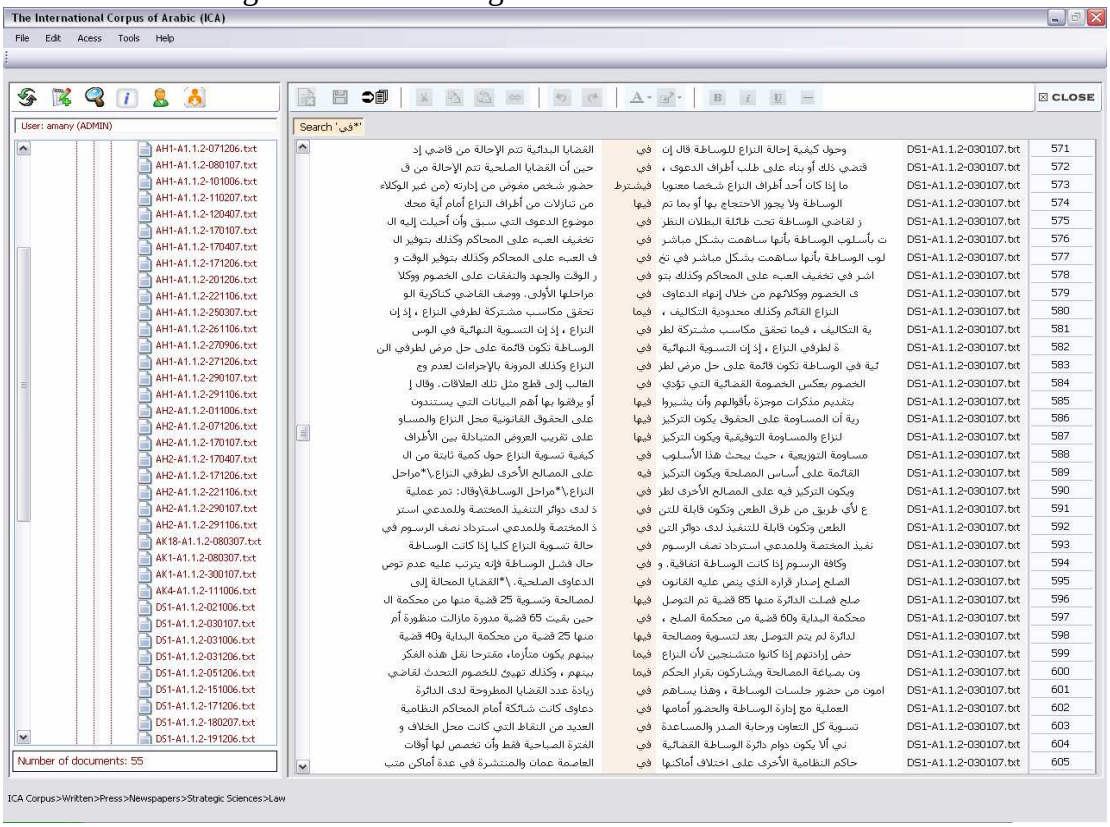

Figure 5: Separate Search Results.

There is another type of options that is the within the context option. The search results of that search option appear in Figure 6:

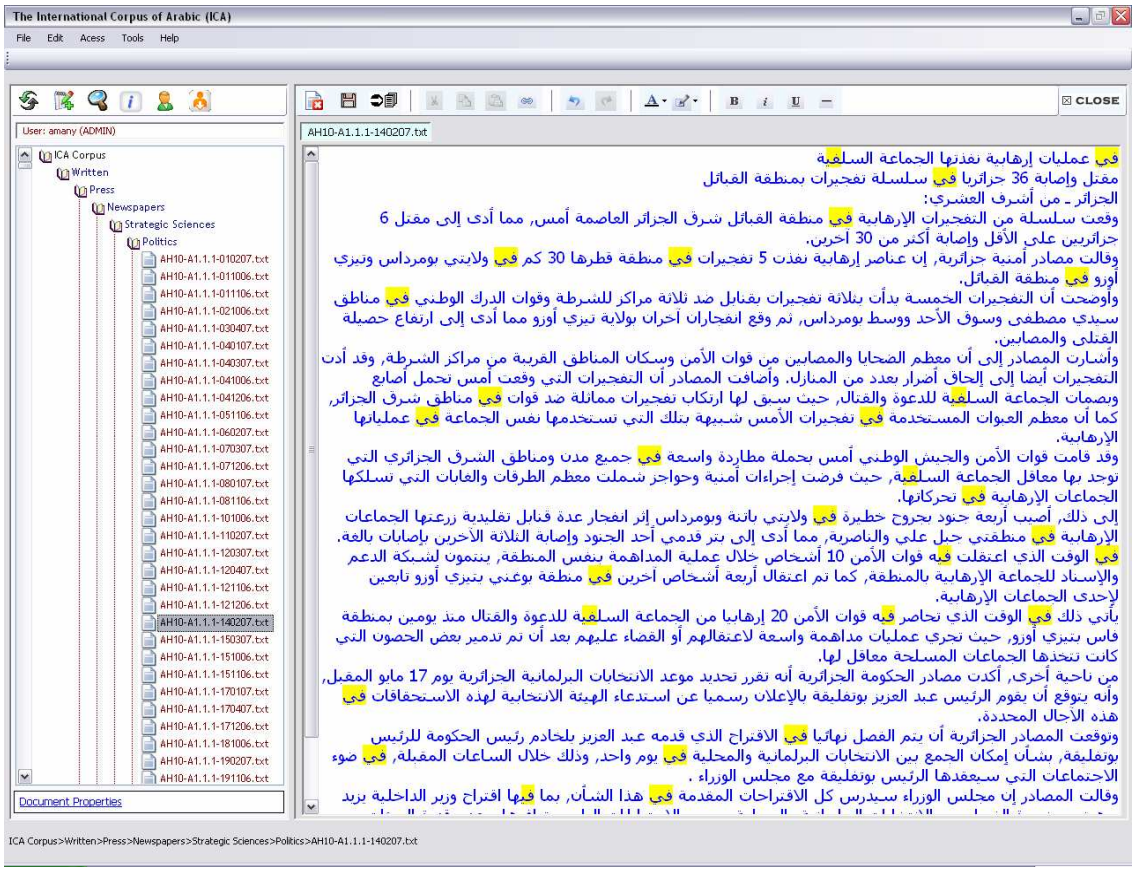

Figure 6: Whole text browsing view

3. **Permissions:** As the system is used on a server, a number of permissions is defined and controlled by the administrator of the system. These permissions are:
   - Read a document and search its contents [R]
   - Copy text from a document [C]
   - Edit a document and its annotations [E]
   - Insert a new document [I]
   - Delete an existing document [D]
   - Control users (add – delete – modify permissions) [U]

6.1 Corpus Map

The ICA software can display the hierarchical design of corpus according to the currently selected node in the tree structure, see Figure 7

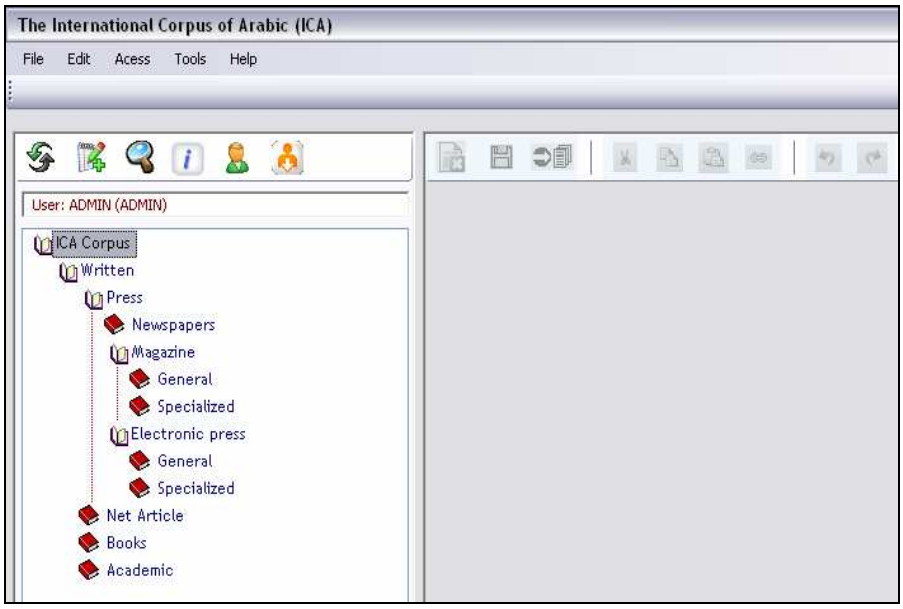

Figure 7:Basic sources and sub-sources

The tree can be expanded to show the basic genres and sub-genres. Figure8, for example, shows the subgenres and sub-sources of the newspaper texts. The pointer can go in depth in the document hierarchy till only a single file in a specific source is selected (figure 9). Once a certain text is selected, it can be displayed in a user-defined font and color. Even texts can be edited if the user has a permission to do this.

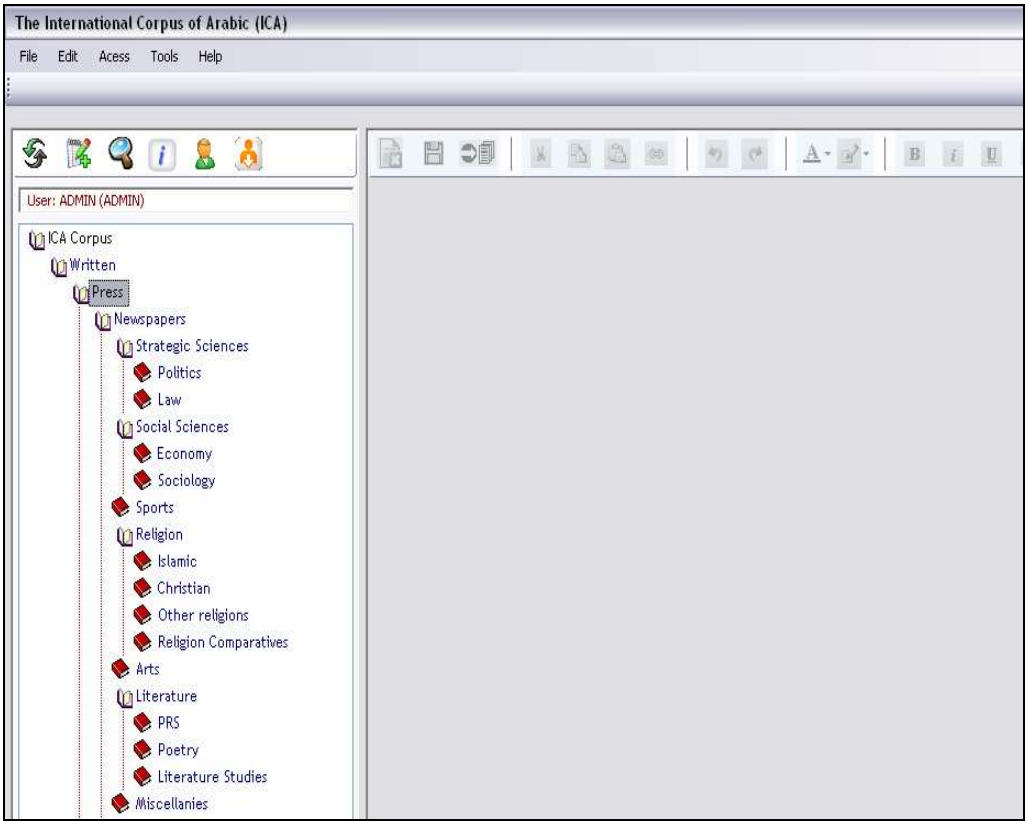

Figure 8:Genres and sub-genres of ICA.

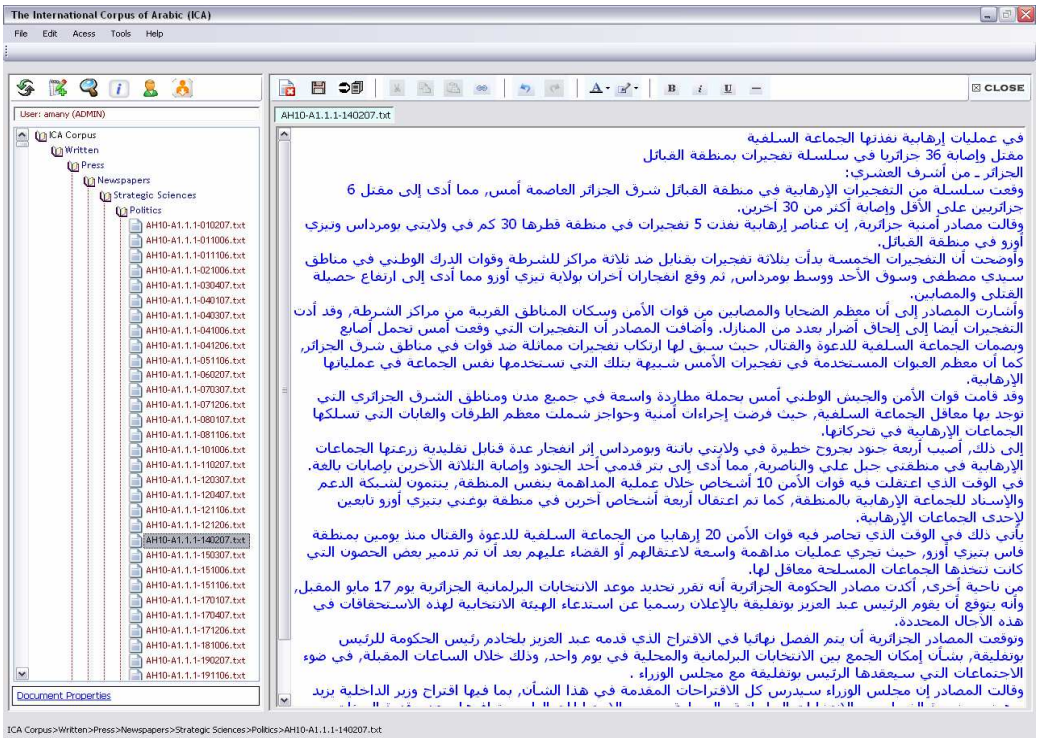

Figure 9: Displaying a given text of the corpus.

7. Conclusions:

It is very important to know that empirical data in language and linguistic researches are essential, due to the issue that this kind of research could not depend on individual cognitive perception. In this view, corpus-based investigations are very helpful; this can be very clear in lexical studies, grammar, semantics, NLP and many other language studies.

Some trials were surveyed for building Arabic corpora, these trials have been presented and evaluated. The paper presented one of the remarkable trials for building a new corpus of Modern Standard Arabic (MSA), the ICA. The current status of the ICA was presented, its design and the preliminary software used in interrogating the corpus. This trial can be considered as one of the successful approached for building a representative corpus of MSA. It is important to realize that the creation of ICA is a "cyclic" process, requiring constant re-evaluation as the corpus is being compiled. It is also important to realize the types of sources, sub-sources, genres and sub-genres to be included in the ICA. Once the process of collecting and computerizing texts is completed, texts will be ready for the final stage of preparation; mark up, and then it will be easy to deal with texts in the analysis stage.

8. References:

Aarts J. (1991), **Intuition-based and observation-based grammars**, Aijmer and Altenburg: 44-62 ,London and New York: Longman.

Aarts J. and Meijs W. (1986), **Corpus Linguistics II**: Rodopi. Amsterdam.

Abdelali A., Cowie J. and Soliman H. (2005), **Building a Modern Standard Arabic Corpus**, Workshop on Computational Modeling of Lexical Acquisition. The Split Meeting. Croatia, 25th to 28th of July 2005.

Aijmer J. and Altenburg B. (1991), **Directions in English corpus linguistics**, Radopi, Amsterdam..

Al Shamsi F. and Guessoum, A. (2006), **A Hidden Markov Model –Based POS Tagger for Arabic**, Conference on the Statistical Analysis of Textual Data, April 19-21, 2006, Besançon / France.

Al-jabri s. (1997) , **Towards the automatic generation of Arabic terminology**, data processing center-kfsc, kingdom of Saudi Arabia.

Al-Sulaiti L. and Atwell E. (2004), **Designing and developing a corpus of Contemporary Arabic**. In Proceedings of the sixth TALC conference. Granada, Spain.

Biber D., Johansson, S., Leech, L., Conrad, S. and Finegan, E. (1999), **The Longman Grammar of Spoken and Written English**, London: Longman.

Biber, Douglas, Conrad S. and Reppen R. (1998), **Corpus Linguistics: Investigating Language Structure and Language Use**, Cambridge University Press.

Collins C. (1987), **COBUILD English Language Dictionary**, London and Glasgow: Collins.

Curme G. (1947), **English Grammar**, New York: Harper and Row.

Duh K., Kirchhoff K. (2005), **POS tagging of dialectal Arabic: A minimally supervised approach**, Association for Computational Linguistics POS Tagging of 2005.

Eldos M. (2002), **Arabic Text Data Mining: A Root Extractor for Dimensionality Reduction**, ACTA Press, Ascientific and Technical Publishing Company.

Eldos T. (2002), **Arabic Text Data Mining: A Root Extractor for Dimensionality Reduction**, ACTA Press, A scientific and technical publishing company.

Goweder A. and Roeck A. (2001), **Assessment of a Significant Arabic Corpus**, Arabic NLP Workshop at ACL/EACL, 2001, Toulouse, France.

Graff D. and Walker K. (2001), **Arabic Newswire Part 1**, Agence France Press.

Guidère M. (2002), **Toward Corpus-Based Machine Translation for Standard Arabic**, Translation Journal, Volume 6, No. 1, January 2002.

Hassan O. and Roukos Y. (2003), **Language Model Based Arabic Word Segmentation**, Annual Meeting of the ACL, In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL.

Heaps H. (1978), Information retrieval, computational & theoretical aspects, Acadimic Press,INC.

Holmes J. (1988), **Doubt and certainty in ESL textbooks, Applied Linguistics**, Oxford University Press.

Holmes J. (1994**), Inferring language change from computer corpora: some methodological problems**, ICAME Journal., No. 18, April 1994.

Johansson S. and Norheim E. (1988), **The subjunctive in British and American English**, ICAME Journal, No. 12, April 1988.

Kennedy G. (1987), **Applied Linguistics**, Cambridge University Press.

Kennedy G. (1987), **Expressing temporal frequency in academic English**, TESOL Quarterly, April 21-25, Miami Beach, Florida, USA.

Kirk J. (1994), **Teaching and language corpora: the Queen's approach**, Teaching and Language Corpora Conference, Lancaster University.

Kjellmer G. (1986), The lesser man: observations on the role of women in modern English writings, Arts and Meijs.

Maamouri M., Bies A. and Buckwalter T.(2004), **The penn Arabic treebank: Building a large-scale annotated arabic corpus**, NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Maamouri M., Bies A.(2004), **Developing an Arabic Treebank: Methods**, Guidelines, Procedures, and Tools, Workshop on Computational Approaches to Arabic Script-based.

Maamouri M., Bies A., Buckwalter T., Mekki W. (2004), **The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus**, NEMLAR Conference on Arabic Language Resources and Tools, 2004.

Maegaard B., Choukri K., Mokbel Ch., Yaseen M. (2005), **Language Technology for Arabic**, NEMLAR, Center for Sprogteknologi, University of Copenhagen, July 2005.

McEnery A. and Wilson and A. (1993), **The role of corpora in computer-assisted language learning**, Computer Assisted Language Learning 6(3).

McEnery T. and Wilson A. (1996), **Corpus linguistics**, Edinburgh University Press.

McEnery T. and Wilson, A. (2001), **Corpus linguistics**, Edinburgh University Press.

Meyer C. (2002), **English corpus linguistics, an introduction**, Cambridge University Press.

Mindt D. (1991), **Syntactic evidence for semantic distinctions in English**, Aijmer and Altenburg.

Mindt D. (1992), **Zeitbezug im Englischen: eine didaktische Grammatik des englischen Futurs**, Tübingen: Gunter Narr.

Oostdijk N. and Haan, P. (1994), **Clause patterns in modern British English: a corpus-based (quantitative) study**, ICAME Journal, No. 18, April 1995.

Otto J. (1949), **A Modern English Grammar on Historical Principles**, Original from the Library of Congress, Vols. 5-7 issued without series title, have imprint: Copenhagen, E. Munksgaard, 1940-49, Digitized Jul 16, 2007.

Quirk R. (1960), **Towards a description of English usage**, Transactions of the Philological Society, Essays on the English Language, Medieval and Modern. London: Longman.

Quirk, Randolph, Greenbaum S., Leech G. and Svartvik J. (1985), **A Comprehensive Grammar of the english Language**, London: Longman.

Renouf A. (1987), **Corpus Development**, in J. Sinclair (ed.), Looking Up, London: Collins ELT, ch.1.

Riesaand J. and Yarowsky D. (2000), **Minimally Supervised Morphological Segmentation with Applications to Machine Translation**, Annual Meeting of the ACL, In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1.

Safadi H., Dakkak O. and Dr. Ghneim N. (2006), **Computational Methods to Vocalize Arabic Texts**, Second Workshop on Internationalizing SSML, Crete, 30-31 May 2006.

Safadi H., Dakkak O., and Ghneim N. (2006), **Computational methods to vocalize arabic texts**, HIAST.

Semmar N., Elkateb-Gara F and Fluhr Ch. (2005) , **Using a Stemmer in a Natural Language Processing system to treat Arabic**, 5th Conference On Language Engineering, Cairo, Egypt, CLE .

Sinclair J. (1987), **Looking Up: An Account of the COBUILD Project**, London: Collins.

Sinclair J. (1992), **Introduction. BBC English Dictionary**, London: Harper Collins.

Tony M. and Wilson A. (2001), **Corpus Linguistics (Second Edition)**, Edinburgh University Press.

Zemanek P. (2001), **Corpus Linguae Arabicae: An Overview**, Proceedings of ACL/EACL 2001 Workshop on Arabic Language.