

Designing and Developing a Corpus of Contemporary Arabic

Latifa Al-Sulaiti
BA, B Ed, MA, PhD

*Submitted in accordance with the requirements for the
degree of Master of Science*

**The University of Leeds
School of Computing**

March, 2004

The candidate confirms that the work submitted is her own and that appropriate credit
has been given where reference has been made to the work of others

Acknowledgments

Studying at the University of Leeds and particularly the School of Computing was the most rewarding experience I have ever had. To me, this is undoubtedly related to the high standard of facilities offered to students, the friendly atmosphere among the research students and above all the excellence of supervision.

For this reason I would like first to express my thanks to the department as a whole and especially to the post-graduate tutor Ray Kwan. Above all, I would like to express my deepest appreciation to my supervisor Eric Atwell - you have been a constant source of encouragement and guidance, and your faith in me is largely responsible for not only completing this thesis but also enjoying working on it.

I must not forget to thank Serge Sharoff, the research fellow at the Centre for Translation Studies, who so generously shared his time and expertise whenever I was confused about corpus encoding. I also would like to thank all my friends in the language research group especially Andy Roberts and Bayan Abu Shawar for their support and help. Being with them made working in the lab very enjoyable.

It is also with great pleasure to record my thanks to the owners of public websites and online magazines for their granting me permission to use their materials for my corpus and to those who participated in my questionnaire. Thank you ALL!

Abstract

Corpora are recognised as an important resource for both teaching and research. Currently, the availability of Arabic corpora is somewhat limited. Appreciating the need for greater research, and enhanced tools and resources, this project has been designed to compile an Arabic corpus, which meets the needs of teachers of Arabic as a foreign language (TAFL), and language engineers.

Early survey research confirmed that existing corpora are particularly restrictive in source-type and genre. It also highlighted a clear demand for a freely accessible corpus of contemporary Arabic, which would include not only Standard Arabic but also samples of colloquial varieties. This perspective reflects current thinking in teaching Arabic as a foreign language. The challenges of sourcing and digitising such material were also recognised from the outset.

In forwarding the project, texts were collected from four main sources - magazines, newspapers, websites and radio - after obtaining appropriate copyright clearance. The texts were then annotated manually using the Unicode editor UNIREL. XML mark-up was used to achieve this deploying a header with the following components: file description, encoding description, and profile description. Other corpora generally consist mainly of raw data, which limits usability.

A corpus of around one million words was compiled covering the major “user” categories. The process of sourcing, categorising, and digitising samples of colloquially spoken Arabic proved difficult. The primary constraints were time, and more importantly, a lack of effective Arabic processing tools. The latter was particularly true in respect of speech-to-text processing where transcription was necessary.

The project highlighted the very real need for the development of dedicated Arabic language processing tools to support corpora development and analysis. Building on this thesis it is hoped that the corpus will be further extended, and will provide a valuable (free) Web resource for many users. In addition, it is anticipated that software engineers will deploy it in the development of dedicated Arabic analysis tools.

Contents

Acknowledgments	i
Abstract.....	ii
Contents	iii
List of abbreviations	v
Chapter 1: Introduction	1
1.1 Corpus Linguistics	1
1.2 English Corpora	2
1.3 Corpora and Teaching.....	3
1.4 Objectives of the Study.....	4
Chapter 2: Arabic Corpus Linguistics.....	5
2.1 Current Arabic Corpora	5
2.2 Arabic Corpus Analysis Tools	15
2.2.1 Arabic Taggers.....	15
2.2.2 Morphological Analysers:.....	17
2.2.3 Optical Character Readers (OCR)	21
2.2.4 Online Arabic Dictionaries	23
2.2.5 Concordancers.....	24
2.2.6 Other tools.....	25
2.3 Summary	25
Chapter 3: Methodology	27
3.1 Method Used.....	27
3.2 Results and Discussion	28
3.3 Forms of Arabic	35
3.4 Teaching Arabic in the 1990's and the Future.....	37
Chapter 4: Corpus Encoding.....	41
4.1 XML and Corpus Encoding	42
4.1.1 Summary	46
4.2 Procedure of the Corpus Encoding	46
4.3 Some Specific Problems	47
4.4 Processing Written Texts and Speech Recordings:	51
4.4.1 Written Texts:	51
4.4.2 Spoken Recordings	51
Chapter 5: Result	54
5.1 The Corpus.....	54
5.2 Copyright Issues.....	59

5.3 Summary	64
Chapter 6: Uses of the Corpus.....	65
6.1 Using Corpora for Language Teaching.....	65
6.2 Using Corpora for Teaching Arabic	68
6.3 Development of a New Concordancer for Arabic	74
Chapter 7: Discussion and Conclusion	77
7.1 Discussion	77
7.2 Contribution	81
7.3 Future Development of the Corpus.....	83
References	84
Appendix I Samples of Existing Corpora	89
Appendix II Buckwalter Transliterating System.....	106
Appendix III Questionnaire.....	107
Appendix IV Corpus Encoding Template.....	111
Appendix V Letters of Copyright	113
Appendix VI Proposal for Extending the Research.....	115

List of abbreviations

AFP	Agence France Presse
ALI	Arabic Language Institute
ANC	American National Corpus
ATC	Air Traffic Control Corpus
BNC	British National Corpus
CAC	Classical Arabic Corpus
CALL	Computer-Assisted Language Learning
CCA	Corpus of Contemporary Arabic
CLARA	Corpus Linguae Arabicae
CRL	Computing Research Laboratory
DATP	Developing Arabic Text Processing systems
DDL	Data Driven Learning
DMT	Developing Machine Translation
DTD	Document Type Definitions
ECA	Egyptian Colloquial Arabic
EATP	Evaluating Arabic Text Processing systems
ELRA	European Language Resource Association
EMT	Evaluating Machine Translation
ESA	Educated Spoken Arabic
GC	Grammar Checkers
GSAC	General Scientific Arabic Corpus
IE	Information Extraction
IRC	Internet Relay Chat
KACST	King Abdulaziz City for Science and Technology
LBCI	International Lebanese Broadcasting Corporation
LDC	Linguistic Data Consortium
LLC	London-Lund Corpus of Spoken British English
LOB	Lancaster-Oslo-Bergen Corpus
MSA	Modern Standard Arabic
MT	Machine Translation
NCCAL	National Council of Culture, Arts and Letters in Kuwait

OCR	Optical Character Reader
OHP	Over Head Projector
POS	Part of Speech
SGML	Standard Generalised Mark-up Language
SMS	Short Message Service
SP	Speech Production
SR	Speech Recognition
STP	Speech to Text Processing
UMIST	University of Manchester's Institute of Science and Technology
TAFLL	Teaching Arabic as a Foreign Language
TEI	Text Encoding Initiative
TSP	Text to Speech Processing
XML	Extensible Mark-up Language

Chapter 1

Introduction

1.1 Corpus Linguistics

Corpus linguistics can be defined as the study of language through the use of large collections of machine-readable texts referred to by the term 'corpora'. Corpus linguistics is not a branch of linguistics but rather a methodology that can be used to study all the aspects of language such as syntax, semantics, pragmatics, speech, and recently in lexicographic studies. The basic corpus methodology was well known in linguistics for a long time but what is different now is the large scale of using corpora in linguistics studies. This is due to the recent explosion in technology especially the massive production of computers and software. The combination of corpora and computers as a means of studying languages changed the way we analyse linguistics phenomena. Linguists are forever curious about different language structures and their functions. In the past many theories and interpretations have been proposed to explain linguistics phenomena. But the scale of data at hand was so small considering the infinity of language. Thus, although results of such traditional studies are accurate, obtaining the result is not very easy. In addition, it was more focussed on investigating language structure rather than on language use (Biber, Conrad and Reppen 1998).

With the advent of corpus linguistics, analysis of language began to go beyond describing the structure of language. It started to analyse the use of structures and investigate the factors that affect our choice of one structure over another. For example, linguists began to look at whether a certain type of structure is used in one type of writing such as science, rather than literature, or whether it is used by women rather than men. Linguistic and non-linguistic factors such as sex, age, period of time, registers, text type, and medium (spoken versus written) are connected to linguistic phenomena and are thoroughly investigated to achieve best result. A linguist's aim is to discover typical linguistic patterns in some defined contexts more than stating their judgments on whether the pattern is correct or incorrect. Such kind of investigation is not easy to do using one's own intuition, but having a big amount of data stored on the computer provides a good resource to carry out this new view of language analysis.

1.2 English Corpora

The first modern electronically readable corpus developed was the Brown Corpus of Standard American English (Kučera and Francis 1967). The corpus consists of one million words of American English texts printed in 1961. For the corpus to be a good representative of the language, the texts were sampled in different proportions from 15 different text categories: press, skills and hobbies, religion, fiction, etc. In the standard of modern corpora, this corpus is considered to be small. However, it is still used in teaching and as a model for the development of other corpora such as the Lancaster-Oslo-Bergen (LOB) (Johansson et al 1986), corpus of British English and the Kolhapur Corpus, a corpus of Indian-English (Shastri 1988). In 1995 another large corpus was developed: the British National Corpus (BNC) (Leech 1993). This corpus consists of 100 million words and it contains both written and spoken material. The ANC (American National Corpus) has recently been under development as a comparable version of the BNC (Ide 2003). In addition to the main varieties of English (British and American), there are other varieties such as Australian (Ahmed and Corbett 1987; Peters 1987), Indian (Shastri 1988), Cameroonian (Tiomajou 1993), and others. Besides these general corpora which can be used for research in various linguistic fields, there are other corpora that are specialised. For example, there are historical corpora such as the Helsinki Corpus of English Texts and also there are corpora that can be used for special purposes like the Air Traffic Control (ATC) corpus¹ and the Trains Spoken Dialogue Corpus². Considering the great value of corpora in research and teaching, many other corpora have been produced for other languages such as French, Spanish, German, Dutch, and many more³

Many recent studies of language adopt a corpus-based approach that can be described as being quantitative as well as qualitative. The quantitative technique can be considered an essential part of corpus-based analysis. When comparing the use of two words or two structures, it is not enough to state their contextual features but to calculate the number of their occurrences or their co-occurrence with other words. Quantitative method requires calculating some statistics to assess the significance of the frequency. This might sound too complicated but since we have the data on the computer we can use some tools to help us get the result we need. Some programs called ‘concordancing programs’ are available for use to obtain any type of analysis ranging from frequencies, to lexical collocations of any type. With

¹ This corpus contains seventy hours of recorded conversation between controllers and aircrafts in three airports in the United States. More information about it is at:

<http://wave ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S14A>.

² Its details can be obtained from <http://www.cs.rochester.edu/research/cisd/resources/trains.html>.

³ See <http://devoted.to/corpora> and <http://clwww.essex.ac.uk/w3c/> for a long list of the existing corpora.

the use of such programs we can control the result and achieve more accuracy by controlling either the linguistic or non-linguistic factors that might affect the occurrence of certain linguistic phenomena.

1.3 Corpora and Teaching

Corpora have long been used for research but it was only in 1992 that some ideas have been proposed to use corpora for teaching (Leech 1997). After the increasing availability of computers, they became an indispensable tool to use for almost anything: shopping, corresponding, looking up a word, searching for information, applying for jobs, etc. Because of its versatility, it was reasonable to investigate its use for teaching. Nowadays corpora resources are used in many universities and institutions to teach languages. It was experimented in the teaching of grammar, translation, vocabulary and many other courses. The many different corpora available made it much easier to use it for almost any educational purpose. And in the domain of learning or teaching a foreign language, it provided a new and exciting venue for learners as well as teachers. Learners are no longer dependent on textbooks but they have a wide range of material to explore and investigate the use of certain structures in different styles and registers. Likewise, teachers can use this resource for either developing teaching materials or for selecting all sorts of teaching activities.

However, it must be pointed out that for the purpose of language/linguistics teaching the present corpora are not completely suitable. On this matter, Leech (1997) states:

'It is a sad fact that the types of corpora which are most easily available for the computer today consist largely of written texts, whereas the types of corpora which would most faithfully reflect the priorities of language learning would contain at least as much spoken material as written material'(1997:17).

This, as he states further, is related to the fact that human beings' experience of language is primarily spoken and secondarily written. As for the data on the computer the reverse is true. For example, spoken material only accounts for 10% of the BNC, with the remaining 90% dedicated to written. Similarly, the Birmingham corpus which has 20,000,000 words contains 10% spoken and 90% written. There are some spoken corpora but they are small. For example, there is the London-Lund Corpus of Spoken British English (LLC) (Svartvik 1990). It consists of 500,000 words and the texts are transcribed and marked with prosodic features. Also there is the Lancaster/IBM Spoken English Corpus which consists of 53,000 words

encoded with several features (Taylor and Knowles 1988). Thus, the spoken corpora are not as large as written corpora and this is related to the high cost of producing them.

The normal trend in corpus linguistics is to produce corpora that represent the standard usage of the language, such was the case of the Brown and LOB corpora. This resulted in producing corpora that do not accurately reflect all the available usage of the language and is connected more with a limited age range (middle-aged to older speakers). However, the BNC contains a wider range of speakers including a corpus for teenage speech: the corpus of London Teenage Language (COLT) (Haslerud and Stenström 1995). Nowadays more corpus builders are directing their attention to wider usages. Several corpora are being constructed based on sources such as emails, Internet relay chat (IRC) and SMS text messages. These sources are produced by teenagers and younger generations. The form of language produced is a spoken language represented in written form. The language exemplifies different characteristics which are represented in the use of slang and coding new meanings to some known lexical items.

1.4 Objectives of the Study

Despite the fact that the state of corpus linguistics in English and other languages has not yet captured the 'real' language by containing more spoken material, it is still more advanced than it is for Arabic. Therefore, it is our purpose in this project to develop an Arabic corpus which will be:

1. based on users' needs.
2. freely accessible.
3. reflects the state of Arabic at the present time.

This thesis consists of six chapters plus a conclusion. Chapter 2 gives an overview of all the available Arabic corpora. It describes their sources, sizes, and contents. Chapter 3 describes the methodology used on which the Arabic corpus was designed. Chapter 4 describes the encoding used for the corpus. Chapter 5 gives background information on the tools available for processing Arabic text. Chapter 6 reports on the corpus developed, its size, and some important issues related to it.

Chapter 2

Arabic Corpus Linguistics

Arabic is an international language, rivalling English in number of mother-tongue speakers (Graddol 1997). The estimated number of native speakers of Arabic is over 200M in addition to over a billion Muslims who use Arabic for practising their religion⁴. Arabic is also one of the official languages in the United Nations and other international organizations. However, relatively little attention has been devoted to Arabic. Although there has been some efforts in Europe, which has resulted in the successful production of Arabic corpora, the progress in this field is still limited. Generally speaking, there is widespread ignorance of Arabic in western universities, due not only to historical and cultural separation but also to the complexity of the Arabic language structure and its unique script. In addition, progress has been impeded by the lack of efficient tools such as tokenisers, taggers, morphological analysers and optical character readers, which are beneficial for developing a corpus.

Section 2.1 describes the currently existing corpora with samples illustrating their formats shown in appendix I. Section 2.2 gives an overview of the different processing tools available for Arabic.

2.1 Current Arabic Corpora

Arab and European scholars who are interested in studying Arabic have developed several corpora, which can be an important research resource since Arabic needs some solid investigation based on large amounts of authentic material. At present, corpus-based research in Arabic lags far behind that of modern European languages. As far as we know, most studies on Arabic up to now have been based on rather limited data.

This section aims to give a brief description of all the Arabic corpora that have been developed and those that are under development.

⁴ As stated at http://www.georgetown.edu/departments/arabic/p_program.htm.

a. Buckwalter Arabic Corpus:

This corpus is developed by the lexicographer Tim Buckwalter. The work on this corpus started in 1986 when electronic Arabic texts did not exist and so the first texts were transcribed from Al-Sharq Al-Awsat newspaper (around 40,000 words), but as the Arabic texts began to appear on the Web, the corpus grew very rapidly to reach between 2.5 and 3 billion words. Its content is derived from public sources on the Web such as newspapers, magazines, news agencies, etc. This corpus was developed for lexicographical purposes to produce word frequency counts, concordances, morphological analysis, and Arabic lexicon. Thus, there is no intention of making it available to the public⁵

b. Leuven Corpus:

This corpus is developed by Mark van Mol at Catholic University Leuven in Belgium (2000). Work on this corpus started in 1990. The purpose of it is to produce a new Arabic-Dutch/Dutch-Arabic learner's dictionary. At present the corpus contains 3M words and is built on three main sources: (a) radio and television news broadcasts from Algeria, Egypt, and Saudi Arabia. More programmes such as interviews, speeches, plays, were added later on. The spoken corpus consists of around 700,000 words. (b) 50 handbooks for learning Arabic which are used in primary schools in nine countries. Such material is chosen because the text is vocalised. (c) Written material from the Internet derived from magazines and newspapers. Because of the difficulty of searching the raw corpus, an encoding system was developed for tagging the corpus to facilitate word searches. The dictionary was completed in 2001. At present, work is still going on to increase the size of this corpus to reach up to 10M words and to produce an electronic version of the dictionary⁶.

c. The Linguistic Data Consortium (LDC):

LDC has several resources for Arabic⁷. Although they are quite large they are far from containing all the types of texts required. It has been pointed out that there is lack of formal speeches, interviews and multi-party discussions in the spoken corpus. In addition, there is lack of literary, scientific and technical genres in the written corpus. The use of the Arabic resources in LDC is restricted to members who pay a certain amount of fees on a yearly basis. The LDC resources consist of:

⁵ See the URL <http://www.qamus.org/wordlist.htm> for more information.

⁶ You can read more about this project at http://www.kuleuven.ac.be/ilt/arabic/index_en.htm.

⁷ See the URL <http://www ldc.upenn.edu/>.

1. The written corpus

There are two corpora: Arabic Newswire Corpus and Arabic Gigaword. The former is a collection of newspaper articles totalling around 80M words. The source of the material comes from the Agence France Presse (AFP), Islamic Republic News Agency, Xinhua News Agency and Ummah Press. The work on this corpus began in 1994 at the University of Pennsylvania. It was for the aim of providing a resource for education and the development of technology. The text is mainly tagged with simple XML to mark the different documents and paragraphs. There is no POS tagging.

Recently the Al-Hayat and An-Nahar corpora have been added causing the written source to increase to around 400M words. This new resource is titled ‘Arabic Gigaword’⁸ (Maamouri and Cieri 2002). The main purpose of this corpus is for applications in natural language processing, language modelling, and information retrieval.

2. The spoken corpus

The source of the material for this corpus was news broadcasts from the radio show, Voice of America. The work on this corpus began in 2000 and it comprises more than 110 broadcasts. However, there is a future plan of obtaining more data from radio and television from other sources, which represent the other regions in the Arabic speaking world. You can see a transcribed sample of this data in appendix I. The text is transcribed using a modified version of Buckwalter’s symbols and is divided into blocks or small sections which are encoded with a string of numbers. The sections can be a sentence, phrase, or even part of sentence or a conjunction. The Arabic text is delimited from any other English or colloquial form by <.....> marking it with syntactic category or stating that it is colloquial. Also other noises such as coughing, crying or silence are marked or described. Below are some examples:

Name of broadcaster

184.36 190.20: havA <N naSir HusaynI> yuHay*Ikum wa na\$raBu (Ca)anbACinA bi Al tafSII bacda havA Al fASil Al mUSlqI

English text

745.50 751.02: <English uh which had been subject to brutal repression in the past but which were not administrated by>

Foreign names (Ugahanda and Rewanda)

⁸ See the URL <http://wave ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T12>.

884.05 887.14: bayna quwAt(in) min <N (Ca)UGandaB> wa (Ca)uxrE min <N rwAndA>

Colloquial

1213.99 1217.13: li mucAhadaBi cAm <Colloquial (Ca)alf wa tisacmiyaB wa (Ci)itnayn wa sabcIn>

3. The Conversational corpus

The conversational corpus is made up of two parts: CALLFRIEND Egyptian-Arabic and CALLHOME Egyptian-Arabic Lexicon. Work on the former began in 1995 to support the development of language identification technology. The corpus comprises 60 unscripted telephone conversations made by Egyptian native speakers to a member in the family or a friend within the United States and Canada. Length of the conversations is between 5 and 30 minutes. The conversations are provided with some useful background information such as speakers' sex, age, education, and call information (channel quality, number of speakers).

The CALLHOME Egyptian-Arabic Lexicon began in 1997 to support large vocabulary speech recognition of speech produced from telephone lines. It contains 120 unscripted telephone conversations between native speakers of Egyptian colloquial Arabic and a family member or a friend. Conversation length is 30 minutes. The basic difference between the CALLHOME and CALLFRIEND is that the calls in CALLHOME originated only within the United States. From this corpus the Egyptian-Arabic Lexicon has been developed. The lexicon consists of 51,202 entries. In addition to the orthographic representation there is also phonological, morphological, stress, source and frequency information for each word. This lexicon represents the first electronic pronunciation dictionary of Egyptian Colloquial Arabic (ECA).

d. Nijmegen Corpus:

This corpus was developed at Nijmegen University, and the team was headed by Jan Hoogland, who was an Arabist (Hoogland 1996). The Nijmegen corpus was designed mainly for producing an Arabic-Dutch/Dutch-Arabic dictionary and was fully subsidised by the Commission for Lexicographic Translation Facilities. Due to the dissatisfaction with the existing Arabic dictionaries, it was decided to create one from scratch instead of translating an existing one. In order to do that, an Arabic corpus has been developed. It contained scanned texts from sources such as Al-Wasat and Al-Arabi magazines, some fiction and non-fiction sources. Also some texts are derived from Al-Hayat and Al-Quds press. Its size is over 2M words. The corpus was compiled in 1995-1996 when the Internet was not such a rich source

for Arabic. At that time Arabic texts on the Internet were mainly images. Therefore, most of the texts of this corpus were scanned.⁹ The dictionary produced from this corpus is now complete.¹⁰

e. CLARA (Corpus Linguae Arabicae)

CLARA was compiled for the development of an Arabic-Czech dictionary. The work on the project began in 1997 in the Institute of Ancient Near Eastern Studies, Charles University, Prague (Czech Republic).

The project was financed by the Grant Agency of the Czech Republic and the Grant Agency of Charles University. It is part of a research plan ‘Czech National Corpus and Corpora of Other Languages’ supported by the Ministry of Education of the Czech Republic.

CLARA is a corpus of written Modern Standard Arabic containing 37M words¹¹ and based on material published in 1975 onwards. It covers Arabic from the Arabian Peninsula, Syria and Egypt. Also, 5% of the samples are from countries such as Tunisia and Morocco. Samples are drawn from different sources: periodicals, books and from other miscellaneous written materials. Some of the materials are scanned while others are obtained through buying commercial texts from the Internet or exchange of scanned texts and gifts. The categories it covers are based on the topic and they are listed table 1.

Table 1: Classification of texts in CLARA

<i>Text categories</i>		<i>Number of words</i>
A	Agriculture	423,897
B	Arts	189,574
C	Fiction	7,766,957
D	Finance	10,394,645
E	Humanities	5,739,692
F	Industry	1,584,438
G	Law	810,671
H	Medicine	757,808
I	Politics	7,505,418
J	Science	1,243,300
K	Sport	939,512
L	Transport	385,680
TOTAL		37,741,592

⁹ The content of the corpus can be found at http://www.let.kun.nl/wba/Content2/1.4.5_Nijmegen_Corpus.htm.

¹⁰ Information about this project is at <http://www.let.kun.nl/wba>.

¹¹ Subsequent to publishing this paper the corpus has been expanded and now its size is 50M words (Zemanek, personal communication).

At this stage of developing the corpus there is no clear decision on the size of the texts in each genre. But the work on balancing the different genres is currently under way (Zemanek 2001).

In connection to the CLARA corpus, several training corpora and databases have been built for different purposes. For example, there is a training corpus with marked morphological boundaries and it consists of 100,000 words, and a training corpus of 15,000 words with annotation of parts of speech. A database has been built for the development of an Arabic-Czech dictionary, but also for works on the annotation of morphological boundaries and POS.

f. Egypt

The Centre for Language and Speech Processing at John Hopkins University developed a machine translation toolkit called Egypt and it runs a parallel corpus of the Qur'an in English and Arabic. This corpus is aligned semi-automatically, and it is available on the Web.¹²

Freeman (2002) points out some problems with this corpus. For example, the length of sentences in the Arabic version exceeds the maximum length which is 41 words. Also in a parallel corpus, texts in both languages should have the same number of sentences. The problem here is that while the English version contains several short sentences, the Arabic version contains long sentences and this creates complications in aligning the two texts. Many of the issues are related to the statistical model within the software developed to handle the parallel corpus.

g. DIINAR Corpus:

This corpus is developed as part of the project DIINAR-MBC which stands for 'Dictionnaire INformatise de l'ARabe, Multilingue et Base sur Corpus. (Multilingual computerised Arabic dictionary). It is a Euro-Mediterranean Project coordinated by J Dichy in Lumiere-Lyon 2 University. The main purpose is to produce a multilingual lexical database in Arabic, English and French using high-level multilingual language resources and natural language processing tools. The items of the lexical databases are based on corpora and are linked with features which cover lexis-grammar relations. This project is completed in 2000. Some of the major achievements of this project are the creation of (a) Arcolex (Arabic Raw Corpora for Lexical purposes). It contains 10M words in Modern Standard Arabic (MSA) (b) Tagged reference

¹² See the URL <http://www.clsp.jhu.edu/ws99/projects/mt/>.

corpus which contains 200,000 words. (c) Indexed marked up corpus of 200,000 words.¹³ It was not possible to obtain a sample of this corpus.

h. ELRA (European Language Resources Association)

This organisation provides two corpora:

1. An-Nahar Newspaper Text Corpus

This corpus contains around 140M words. The articles, which are in Arabic (Lebanon) from 1995 to 2000, are stored as HTML files on CD-ROM media. Each year contains 45,000 articles and 24M words. Each article includes information such as title, newspaper's name, date (English calendar), country, type, page, column number, etc.

2. Al-Hayat Corpus (Dataset)

This corpus was joint project between the University of Essex and the Open University. The corpus contains 18,639,264 tokens in 42,591 articles, organised in 7 categories: general, cars, computers, news, economics, science, and sport. This corpus has been marked-up for Language Engineering and Information Retrieval purposes. Thus, numbers, special characters and punctuation have been removed. The corpus contains some SGML markup and every text in the corpus is separated by a line from the other and encoded with information such as date of Arabic and English calendar, issue number, page number, title of the text and topic which is supplied at the end of the text. See appendix I for a sample of the corpus exhibiting this encoding.

i. English/Arabic Parallel Corpus

This corpus was developed at the University of Kuwait and funded by the Kuwait Foundation for the Advancement of Sciences (Al-Ajmi 2003). Its aim is to improve bilingual dictionaries and to develop collocational dictionaries. It is also to be used in teaching and research. The total size of the corpus is 3M words, which is not sufficiently large to be representative of the language. Therefore, it is considered to be a prototype and there is a plan in the future to increase its size. The texts of the corpus are derived from the series 'A'laam Al-Ma'rifa' (the World of Knowledge) which is published by the National Council of Culture, Arts and Letters in Kuwait (NCCAL). Some of these monthly books are translations from English and so there was no need to translate the texts but rather to locate the English versions at the NCCAL's

¹³ More details about this project can be obtained from http://sites.univ-lyon2.fr/langues_promodiinar/Accueil.htm.

library. The texts obtained are those that are only published in the 90's so that the focus is on the current usage of the language. The text categories that were included in the corpus are history, economy, arts, literature and general science. However, there is no exact figure for the number of words in each category except that 25% of the words belong to history and economy. The corpus is available in CD-ROM and on the Web but it can only be accessed via a password as its use is restricted to users in the University of Kuwait. The source and target texts were scanned using OCR software and saved as XML files. They were aligned at the sentence level using an aligner program and the search tool is 'Al-Idrisi' which is developed by Sakhr software. See appendix I for a sample of this corpus.

j. Multilingual Corpus

This multilingual corpus was developed by Sattar Izwaini (2003) at the University of Manchester Institute of Science and Technology. It is a corpus of information technology in English and two translational corpora: Arabic and Swedish. Its purpose is to investigate how the lexis of information technology and lexical collocations are translated into Arabic and Swedish. The size of the three languages is not equal and this is related to the availability of texts for each language and whether it is easy to get permission from copyright holders. Therefore, the English corpus contains 7M words. The Arabic corpus contains 1M words and the Swedish corpus contains 2.7M words. The texts are mainly collected from books, research papers, and from websites and online help for computer systems and software. This corpus will not be available for public use as the copyright permission was obtained mainly for the researcher's investigations.

k. 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. The material is derived from the Kuwaiti magazine site 'Science and Technology'.¹⁴

Part of this corpus (1M) has been tagged. Al-Muhanna reported that his training corpus contains 100,000 tokens and the accuracy of his tagging is 92%. Below is a small sample of this tagged corpus.

¹⁴ It is found at <http://www.kisr.edu.kw/science/>.

/JJMS /AT /NP /NP /JJMS /AT /NNM /AT /VB
 /IN /NNSF /AT /NNM /PLNMF /NNF /NNF /IN
 /AT /PP\$3FS /NNF /IN /PPSF3 /CC /NNF /AT /NNM
 /JJFS ./ . /AT /JJFS

However, this corpus is only created for the purpose of researching a specific topic and the developer is not intending to make it available for public at this stage (Al-Muhanna 2003).

I. Classical Arabic Corpus (CAC)

Established by Abdel-Hamid Elewa at the University of Manchester Institute of Science and Technology, this corpus 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 two websites.¹⁵ It contains 5M words. This corpus is not tagged at this stage as it was mainly developed for the purpose of lexical analysis. The corpus can be considered to be representative of that period because it contains all the texts that are available. The main division of the corpus is intended to be between fiction and non-fiction. However, since fiction represents only 11% which is due to unavailability of fictional material for this period, the text types are divided into four genres: thought and belief, literature, linguistics, and science. The table below gives a list of these genres with their sub classifications:

Table 2: Classification of texts in CAC

	<i>Text categories</i>	<i>Number of words</i>
A	The holy Qur'an	88,622
B	Biography	393,933
C	Fiction	579,223
D	Hadith	683,970
E	Linguistics	404,080
F	Philosophy	478,141
G	Poetry	69,385
H	Proverbs	362,054
I	Science	903,205
J	Theology	1,037,387
	TOTAL	5,000,000

The main features of this corpus are that the texts are based on whole books and the focus is only on the main texts. So information such as bibliography, footnotes, and tables are omitted from the corpus. Also the books are written by different authors so that the corpus provides a variety of styles (Elewa 2004).

¹⁵ They are www.muhammadith.org and www.alwaraq.com.

m. SOTETEL-Corpus

Société Tunisienne d'Entreprises des Télécommunications (SOTETEL-IT) developed a corpus of 8M words. The corpus contains a variety of genres such as literature, academic and journalistic writings. Its main purpose is for lexicographic research (Nikkhou & Choukri 2004).

Table 3 below summarises the result of the corpora survey in the order they commenced¹⁶.

Table 3: Classification of Arabic untagged corpora

<i>Name of Corpus</i>	<i>Source</i>	<i>Medium</i>	<i>Size</i>	<i>Purpose</i>	<i>Material</i>
Buckwalter Arabic Corpus 1986-2003	Tim Buckwalter	Written	2.5 to 3 billion words	Lexicography	Public resources on the Web
Leuven Corpus (1990-2004)	Catholic University Leuven, Belgium	Written and spoken	3M words (spoken: 700,000)	Arabic-Dutch /Dutch-Arabic learner's dictionary	Internet sources, radio & TV, primary school books
Arabic Newswire Corpus (1994)	University of Pennsylvania LDC	Written	80M words	Education and the development of technology	Agence France Presse, Xinhua News Agency, and Umma Press
CALLFRIEND Corpus (1995)	University of Pennsylvania LDC	Conversational	60 telephone conversations	Development of language identification technology	Egyptian native speakers
Nijmegen Corpus (1996)	Nijmegen University	Written	Over 2M words	Arabic-Dutch / Dutch-Arabic dictionary	Magazines and fiction
CALLHOME Corpus (1997)	University of Pennsylvania LDC	Conversational	120 telephone conversations	Speech recognition produced from telephone lines	Egyptian native speakers
CLARA (1997)	Charles University, Prague	Written	50M words	Lexicographic purposes	Periodicals, books, internet sources from 1975-present
Egypt (1999)	John Hopkins University	Written	Unknown	MT	A parallel corpus of the Qur'an in English and Arabic
Broadcast News Speech (2000)	University of Pennsylvania LDC	Spoken	More than 110 broadcasts	Speech recognition	News broadcast from the radio of voice of America.
DIINAR Corpus (2000)	Nijmegen Univ., SOTETEL-IT, co-ordination of Lyon2 Univ	Written	10M words	Lexicography, general research, NLP	Unknown
An-Nahar Corpus (2001)	ELRA	Written	140M words	General research	An-Nahar newspaper (Lebanon)
Al-Hayat Corpus (2002)	ELRA	Written	18.6M words	Language Engineering and Information	Al-Hayat newspaper (Lebanon)

¹⁶ Recently an exhaustive survey of the Arabic language resources (totalling 90) has been published by Nemlar (Network for Euro-Mediterranean Language Resources): <http://www.nemlar.org>.

				Retrieval	
Arabic Gigaword (2002)	University of Pennsylvania LDC	Written	Around 400M	Natural language processing, information retrieval, language modelling	Agence France Presse, Al-Hayat news agency, An-Nahar news agency, Xinhua news agency
E-A Parallel Corpus (2003)	University of Kuwait	Written	3M words	Teaching translation & lexicography	Publications from Kuwait National Council
General Scientific Arabic Corpus (2004)	UMIST, UK	Written	1.6M words	Investigating Arabic compounds	http://www.kisr.edu.kw/science/
Classical Arabic Corpus (CAC) (2004)	UMIST, UK	Written	5M words	Lexical analysis research	www.muhammadith.org and www.alwaraq.com
Multilingual Corpus 2004	UMIST, UK	Written	10.7M words (Arabic 1M)	Translation	IT-specialized websites
SOTETEL Corpus	SOTETEL-IT, Tunisia	Written	8M words	Lexicography	Literature, academic and journalistic material

2.2 Arabic Corpus Analysis Tools

In order to develop a corpus and make it more useful for teaching and research purposes, a number of tools must be available. They include optical character readers, morphological analysers, on-line dictionaries, concordancers and taggers. This section gives an overview of the tools that have been developed so far to use with Arabic.

2.2.1 Arabic Taggers

1. Khoja's APT tagger

The APT (Arabic Part-of-Speech Tagger) tagger works directly on Arabic text and was developed using a combination of both statistical and rule-based techniques. The tagset which consists of 177 tags is based on the description of traditional Arabic grammar and thus it is divided into three major classes: nouns, verbs, and particles. Adverbs and prepositions are treated as subcategories of the main classes. The tagsets are assigned to complete words. That is, the word with all its affixes (Khoja 2001, Khoja et al 2003, Khoja 2003).

A stemmer was implemented with a rule-based tagger and a statistical tagger, and was designed to handle the different stages in tagging. The function of the stemmer is to remove all of the words' affixes to produce the stem or root. Since the grammatical category of the word is recognised from its affix. Therefore, the affixes are used to determine the tag of the word. Tests of the stemmer over Arabic words show that it achieves a success rate of 97% using a dictionary of 4,748 trilateral and quadrilateral roots. The ambiguous words are handled by the statistical tagger. Two probabilities are used: lexical probability, and contextual

probability. The statistical tagger achieves an accuracy of around 90% when disambiguating ambiguous words. But for achieving a better result manual tagging is used and pre-processing component needs to be added to handle some mistakes such as the hamza and the glottal stop and the dots under the *y*. The overall accuracy of this tagger has been recorded to be 86%. So far this tagger has not been made available for public use. A sample of a tagged text is given below:

NCSgMND_الملك NCDuMGD_الشرفين NCDuMAD_الحرمين NCSgMNI_خادم VPSg3M_بعث
 NCSgFNI_برقية NP_سعود R_ NCSgMAD_العزیز NCSgMAI_عبد NCSgMNI_بن NP_فهد
 RF_كواسنيفيسكي RF_الكسندر NCSgMGD_الرئيس NCSgFGI_فخامة PPr_الى NCSgFGI_تهنئة
 NCSgFGI_جمهورية NCSgMNI_رئيس

(From Al-Jazirah newspaper dated 1/1/1998, Khoja 2002, p.7)

2. Freeman's Arabic version of Brill Tagger

Another part-of-speech tagger developed for Arabic is based on Brill's tagger (Freeman, 2001, 2002). His tagset has 146 tags and is loosely based on the Brown corpus tagset for English. Since this tagset is designed for Indo-European languages, naturally it includes tags for categories that traditional Arabic grammar does not recognise or lacks some categories that Arabic has such as the dual and feminine nouns and adjective. His tagger works on Roman characters and is applied to lexemes rather than complete words.

Thus, the first step is to transliterate the Arabic text into Roman characters using the system devised by Buckwalter and Beesely (2001). The second step is to analyse each grapheme into its lexemes, which are stems, and affixes. And the final step is applying the tags. For example, the word "فسيكتبونها"/fasayaktubuunaha/ 'and so they will (3rd pp) write (Pl) it (fem)' can be analysed into 6 lexemes and each receives a tag:

Fa-sa-ya-ktb-uwna-haA ---→ fa(CC) sa(FUT) ya(PPI3) ktb(VB) uwna(PLURAL)
 haA(PP\$3FS)

Every lexeme has a single lexical entry in the lexicon. The main purpose is to tag a 50,000 word corpus to use for the training phase of Brill's tagger. So far a corpus of 20,000 words of Modern Standard Arabic has been segmented and tagged and also a relatively small corpus of spoken Yemeni Arabic (affixes specific to the dialect are added to the lexicon). This tagger has been available from the Arabic mailing list (Arabic-l@byu.edu) for trial.

Below is a sample of tagged text using Freeman's tagger (Freeman personal communication, 2002).

ya/PPI3 qym/VB Aal/AT n\$AT/NNM Aal/AT AjtmAEy/JJMS mhrjAn/NNM aA/DUAL
fy/IN mrkz/NN \$bAb/NNSM Aal/AT dwHp/NN Dmn/IN fa/CC EAlyp/JJFS aAt/PLNMF
brnAmj/NN na/PPI1P \$AT/VB hu/PPO3MS Aal/AT AjtmAEy/JJMS Aal/AT trfyhy/NN
Aal/AT hAdf/JJMS

3. LDC Tagger

LDC is developing a POS tagger for Arabic. This tagger is based on the automatic annotation output produced by Tim Buckwalter's morphological analyser of a corpus consisting of 734 files from the Agence France Presse. This tagger is developed by Maeda Kazuaki and Hubert Jin, and it runs on Sun workstations using a Python program. 7 annotators who are native speakers of Arabic participated in the project. At this stage the tagger has some problems. It has been reported (Maamouri and Cieri 2002) that the most frequent problems encountered with this tagger are: the absence of non-Arabic proper names, place names, and company names, some foreign names are identified as Arabic words, e.g. *Minh* as *minhu*, not identifying the correct short vowels especially in the case of passive voice. Further work on this tagger is still going on to improve it and overcome all the problems.

2.2.2 Morphological Analysers:

The main function of a morphological analyser is to identify the roots of words and the morpho-syntactic information related to them. Therefore, having a morphological analyser is important in processing natural language. For Arabic there is a need for a powerful and fast system to cope with texts from different domains with a variable document structure and written in different writing styles. Moreover, when processing Arabic texts from the Internet we are faced with unvoiced texts that are more confusing. Handling words not using diacritics is ambiguous since usually they can have more than a single analysis.

The last decade has seen the development of several morphological analysers, such as the Xerox morphological analyser (Beesley 2001, 2003), Buckwalter morphological analyser (2002), Sakhr morphological analyser, Darwish morphological analyser (2002) and an analyser by Berri, Zidoum, and Atif (2001). In this part brief descriptions of some of the morphological analysers which have been developed will be presented.

1. Beesley's Xerox Arabic Morphological analyser and generator

Beesley (2001, 2003) developed an Arabic morphological analyser and generator using finite state techniques.¹⁷ The purpose of this morphological analyser is to use it as a teaching aid and as a component in larger natural language processing systems. It is a two-level morphological analysis: one level is for roots and patterns and the other level is for affixes, prefixes, enclitics, and some forms such as prepositions, definite article, conjunctions which are normally attached to words as prefixes. By certain alternation rules the two levels are combined to produce all the acceptable occurrences of the words.

The morphological analyser uses an Arabic dictionary containing 4930 roots that are combined with patterns. The various combinations of prefixes and suffixes with the stems produce over 77,800 tokens. This tool analyses words that include full diacritics, partial diacritics, or no diacritics. When a user inputs a word, the morphological analyser gives all the occurrences with the short vowels. It gives also the root morpheme and the pattern morpheme. For example, for the word *drs* you can get 10 occurrences. And for every occurrence you get the root *drs* meaning 'study' and the pattern represented in the form of CvCvCv where the v is replaced with the actual vowels and all the grammatical features of the word. It must be pointed out that this tool analyses words in isolation. One of the things we observed about the organizations of the solutions given to an input is that they are not ordered according to their frequency of use. For example, we have input the word *katab*, which has the frequent meaning of either 'wrote', or 'a book', we received 15 solutions but the frequent meaning came as number 9 and 11. If this tool is to be used as a teaching aid, it might be sensible to organise the solutions according to their frequency of use. This system is available free for computational linguists online.¹⁸ However, it must be pointed out that users might encounter problems using it depending on the browsers they are using. It works well with Mozilla browser but not with Internet Explorer. The reason for that is that the Arabic demo (Keyboard entry page) includes a large Java applet which needs to be executed by the Java Virtual Machine (JVM). In some cases, versions of the JVM have bugs that prevent the running of the applet.

¹⁷ It can be accessed at <http://www.xrce.xerox.com/competencies/content-analysis/arabic/>.

¹⁸ See <http://www.xrce.xerox.com/research/mltt/arabic>.

2. Buckwalter Arabic Morphological Analyser Version 1.0

This morphological analyser is freely accessible online¹⁹. It is used by Linguistic Data Consortium (LDC) for POS tagging of Arabic texts. It contains over 77,800 stem entries which represent 45,000 lexical items (Maamouri and Cieri 2002). The parser output uses a transliteration system in which each single symbol corresponds to one unit of Arabic script grapheme (See appendix II). However, the use of this system has been criticised (Beesley 2003) on the grounds that it is not a ‘recognized standard’ of any type, it is not easy to mix Arabic and Roman text in the same document, and it fails to represent the Arabic punctuation because they are used to represent letters.

3. Berri, Zidoum, and Atif Morphological analyser

Berri, Zidoum and Atif (2001) developed another morphological analyser. The main characteristic of this tool is that it treats unvoiced texts derived from Internet. This is an advantage over the other morphological analysers as they are almost all based on entries from dictionaries. In order to implement this tool the ‘contextual exploration method’ is used. The essence of this method is that it identifies the token and its contextual features then it looks for the suitable affixes to be associated with it.

The system of the morphological analyser developed here consists of three main components: a rule knowledge base which has the regular and irregular morphological rules of the Arabic grammar, a set of word lists containing the exceptions handled by the irregular rules, and a matching algorithm that matches the tokens to the rules. So the text input from the Internet source goes through a number of modules. First, all the useless parts are removed from the document and the HTML tags are identified and coded. The tokeniser identifies all tokens and sentences. Then a module builds an object-oriented representation of the text that highlights all the basic relationships between the different constituents of the document, namely the sentence, the paragraph, the section and the title. Then finally the morphological analyser finds all word root forms and links the morpho-syntactic information to the token. For example, every token has information such as: name, value, root, category and other grammatical features, rule applied, reference to the sentence in which it contains, order of the token in the sentence, position of the token in the text, format (bold, italics, etc.). The way the system works is that the matching algorithm attempts to match between the affixes of the token with a regular rule. If it does not succeed, it applies an exception rule by searching in

¹⁹ See <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49>.

the exception lists. So far only regular rules have been implemented and the rest of the work is still under progress. The difference between this morphological analyser and that of Beesley is that it gives the analysis of tokens based on their linguistic contexts.

4. Sakhr's Morphological analyser

Sakhr Company also produced a morphological analyser, which is referred to as a Multi-Mode Morphological Processor (MMMP). This program covers modern and classical Arabic and it identifies the base form by removing all the affixes and it gives the morphological pattern. Decomposing Arabic words into their morphological primitives is a basic requirement for machine processes such as: full text indexing, search, dictionary organisation and look-up as word spelling and grammatical checking.

5. Darwish's morphological analyser

Darwish (2002) developed a morphological analyser for Arabic which he described as being 'shallow' because it only gives the possible roots of any given Arabic word. Therefore it can be considered as a stemmer. It consists of two modules:

a. 'Build-Model' module: it uses a list of word-root pairs as inputs to derive a list of prefixes and suffixes, to form stem templates and to compute the likelihood that a prefix, a suffix or a template would appear. The word list (279,606 words) is constructed automatically using the Arabic morphological analyser ALPNET.

b. 'Detect Root' module: calculates the probabilities of stems, suffixes and templates as occurring in one combination. The problem is that it over-generates and produces nonexistent words. This problem is solved by matching the words with a dictionary list and checking them manually. The recorded speed of this stemmer is that it derives the roots of 40,000 words per minute on a Pentium class machine.

Although this stemmer is freely available online²⁰, we were unable to execute once downloaded.

²⁰ See <http://www.glue.umd.edu/~dlrg/clir/arabic.html>.

6. Other analysers

There are other morphological analysers produced by companies such as the one produced by Cimos Company (Cimos 2004), and the Engineering Company for computer systems development in Egypt (RDI) (Nikkhou & Choukri 2004).

2.2.3 Optical Character Readers (OCR)

One of the important tools for developing a corpus is having a good OCR for scanning texts that are not available in an electronic form. The quantity of Arabic texts available in digital format does not cover all kinds of materials especially novels and well-known books. This section presents some information about the Arabic OCR software available on the market, and also presents some evaluations which have been reported in the literature. The purpose here is to report on the progress in the technology for Arabic processing and to compare the performance of Arabic OCR's that are available on the market for the purpose of identifying the one with high-accuracy and more useful features. There are a handful studies that have attempted to evaluate OCR's. Some of the products that are available earlier on and which have been evaluated are: TextPert v3.7 Arabic and Al-Qari' Al-'Ali²¹ v1.1 (Bell and Zemanek 1994), ICRA v4.0 and Al-Qari' Al-'Ali v1.1 (Hoogland 1996), Arabic OmniPage v2.0 and Sakhr Automatic Reader v3.01 (Kanungo et al 1998).

1. TextPert v3.7

It is produced by CTA, Inc. and runs on the Macintosh Arabic system. It has been reported (Bell and Zemanek 1994) that this program is easy to handle, but it has its limitations, as it does not come with many trained fonts. When scanning good texts, the result is acceptable but in texts with complicated fonts the program does not function properly.

2. Al-Qari' Al-'Ali v1.1

This program has been tested by Bell and Zemanek (1994) who reported that despite some problems such as having breaks and spaces between letters within words, the result was impressive. One of the main problems that has been discovered was that it was slow when it comes to training new fonts. Therefore, in order to improve it, it has been suggested that it should be provided with more pre-trained fonts.

²¹ It is also known as Automatic Reader

3. Al-Qari' Al-'Ali v1.1 and ICRA v4.0

In another study by Hoogland (1996) compares the performance of Al-Qari' Al-'Ali with ICRA. He points out that both programs can recognise image files of various formats, recognise fonts that have to be trained, and save the recognition results in text-files in various code-pages. However, Al-Qari' Al-'Ali has an advantage over ICRA in the following features:

- a. Provide more interactive training of fonts in a user friendly way.
- b. Recognition during training where the program only stops at characters that have not been trained.
- c. Spell-check option in relation to image file and text file.
- d. In the spell-checking mode the user can switch to the training mode in order to add a character or combination of characters.
- e. Various possibilities for framing by including or excluding text blocks for recognition.
- f. Facilities for storing user-defined combination of characters up to maximum of 4 combinations.
- g. Powerful batch mode processing. In batch mode the program recognise a several image files using a specified font. While this is being processed, the user can carry out other activities.

This program has been tested on several text types such as magazines and literary tests and found that the recognition accuracy was approximately 97-99%. One of the difficulties with this program is that it does not recognise punctuation, Latin characters, and numbers. Also there is difficulty in processing newspaper texts as some of the letters, which are not supposed to be connected, are connected to the following characters such as the case of *w*.

4. Arabic OmniPage 2.0 vs. Sakhr Automatic Reader 3.01

In a study by Kanungo et al (1998), two OCR's: Sakhr Automatic Reader 3.01 and OmniPage 2.0 have been compared using the paired model approach. They have shown that on the 300dpi SAIC dataset Sakhr had higher accuracy than Omnipage but OmniPage 2.0 has a better precision. The average page accuracy rate of Sakhr is 90.33% while that of OmniPage is 86.89%. The average page accuracy of Sakhr is $3.44 \pm 1.13\%$ higher than that of OmniPage. But at 300 dpi OmniPage has 0.99 ± 0.47 higher precision than Sakhr. The accuracy of Sakhr drops when the image resolution is increased beyond 300 dpi.

The latest version of OmniPage is v14. According to the information obtained from the Internet, it seems to work on languages such as English and German and it is considered to be the best OCR, but it does not handle Arabic.

5. Sakhr's OCR's

In addition to Sakhr Automatic Reader 3.01, there are several new versions of OCR's, which are produced by the same company:

- a. Universal OCR: runs on Microsoft Windows. This OCR gives good results if it is trained for the language or fonts.
- b. Sakhr's Al-Qari' Al-'Ali office version: it is similar to the professional version except that it does not have a spell checker and it has only a single mode.
- c. Sakhr's Al-Qari' Al-'Ali v7.0 which comes in two professional versions: gold edition and platinum edition²².

The new professional version has been reported by those who used it to be the best in the market (Freeman, personal communication, 2003). Some of the features it has are:

- High recognition accuracy which is around 99% for good images and 97% for bad images.
- Supports mixing between Arabic and English text.
- Contains Arabic English spell checker.
- You can recognise or skip diacritics.
- Support 4 different types of batch mode: scan and save, scan and recognise, load and recognise, scan and save then load and recognise.
- Supports tables, columns, images, page size detection, and paragraph indentation.

2.2.4 Online Arabic Dictionaries

There are several English and European online dictionaries such as Cambridge dictionary, Oxford dictionary, Webster's dictionary, Oxford Spanish dictionary, and many others. However, in Arabic the number is still limited and cannot be sure if the content is built on a corpus or derived from the hard copies available on the market. There are some free online ones such as Ajeeb²³. It is a bi-directional dictionary but what is on the Internet is not the complete reference work. There are also other online dictionaries such as Ectaco and Lisan

²² See <http://www.aramedia.com/ocroptions.htm>.

²³ It is produced by Sakhr Company. See <http://dictionary.ajeab.com>.

Al-Arab (available on CD and distributed by Markaz Al-Turath li Abhaath Al-Haasib Al-Aaly), Al-Mawrid, and Al-Misbar (produced by ATA Software Technology). Cimos Company which is based in France also produced four online dictionaries (English to Arabic, French to Arabic, Arabic to French, Arabic to English) such as Ad-Dalel general dictionary which contains over 150,000 basic entries and Ad-Dalel specialised dictionary covering areas such as computing, science, business, etc., and it contains 30,000 basic entries²⁴. However, all of these are not free to use.

In a study conducted recently by Al-Ajmi (2002) to test the efficiency of English-Arabic dictionaries, he examined the errors and recorded the difficulties made by 46 English majors at Kuwait Univerisy when consulting two dictionaries: Al-Mawrid and Oxford English–Arabic Reader’s Dictionary (ORD). He found out that the successful lookups in Al-Mawrid were 67.8%, the unsuccessful lookups 29.8% and 2.4% were not listed. A similar result was found for Oxford: the successful lookups 68.5% and the unsuccessful lookups 28.4% but with a higher rate of unlisted items (3%) which is due to its small number of entries. Al-Ajmi concluded that there is a need for urgent improvement in the design of bilingual dictionaries in the Arab world and one source of solution is a large parallel or comparable corpus that would provide natural data and bring dictionaries to the satisfaction of the users.

2.2.5 Concordancers

Concordancers are tools used to search a corpus for any kind of linguistics information such as meaning of lexical words or phrases and investigating certain grammatical structures. There are several concordancers such as Wordsmith which was developed by Mike Scott²⁵, Monoconc²⁶ and Paraconc²⁷ which were designed by Michael Barlow, and they all work perfectly well on languages with Roman letters such as English and other European languages. However, there are few concordancers that handle the unique scripts of Arabic. At present researchers use Monoconc which works on Arabic only in Microsoft Arabic Windows. However, the result of the concordance is not displayed correctly (See chapter 6 for more detailed information). Boualem, Leisher and Ogden (1996) developed a concordancer called xconcord which handles Arabic better than Monoconc does. This concordancer is developed at the Computing Research Laboratory (CRL) at New Mexico State University. One of its characteristics is that it uses Unicode; it supports 17 languages including Arabic, it displays Arabic concordance file correctly from right to left and provides several

²⁴ Information obtained from <http://www.cimos.com/index.asp?src=fiche#13>

²⁵ A trial version can be downloaded from: <http://www.lexically.net/wordsmith/>.

²⁶ Information about this concordancer is available at: <http://www.ruf.rice.edu/~barlow/mc.html>

²⁷ Information about this program can be obtained from: <http://www.ruf.rice.edu/~barlow/pc.html>

sophisticated searches. However, its main limitation is that it works only on Sun Solaris operating system. Also, the work on this tool has never been developed further.²⁸ However, very recently Roberts (2004) developed a new concordancer for Arabic which can be referred as aConCorde v0.4. It not only displays Arabic correctly but it can be used on many platform systems such as non-Arabic Windows, Linux, Apple Mac and others (More detailed information about this concordancer will be provided in Chapter 6).

2.2.6 Other tools

There are other tools which are produced by different institutions and companies. Below is a brief list of them (Nikkhou and Choukri 2004):

- Arabic Pen: Arabic Handwriting Recognition developed by Arabic Textware in Jordan.
- Spell checker: underdevelopment in Amman University
- Arabdiac: an automatic diacritiser produced by the engineering company for computer systems development in Egypt (RDI). Another one is produced by Cimos Company in France.
- Arabic grammar checker: produced by Sakhr company
- Entity extractors, fact extraction systems, cross lingual information retrieval, text mining all produced by Xerox.
- Speech recognition and speech synthesis: produced by Sakhr Company and KACST in Saudi Arabia.

2.3 Summary

From the previous review it can be said that most systems of morphological analysers that were found published references on, were either unavailable or very difficult to use. Furthermore, there seem to be no agreed standards on what the output of an Arabic morphological analyser should be: adding vowels to unvowelled text, and/or finding roots and affixes, and/or adding morphosyntactic features, and/or adding part-of-speech tags; and if the latter, there is no agreement on appropriate POS-tagset equivalent to European EAGLES standard (Leech and Wilson 1999).

As for the Arabic corpora, they are not readily accessible to the public except the corpus 'Egypt'. Several of the corpora are archived with corpus repositories (the Linguistic Data Consortium (LDC) in Pennsylvania and the European Language Resource Association

²⁸ It can be download it from <http://crl.nmsu.edu/Tools/Software/>.

(ELRA) in Paris) from which they can be purchased by academic or industrial research organisations; however, this does not make them readily accessible to most TAFL researchers and practitioners. Others are generally collected for a specific research project rather than as a general resource. In contrast, some English language corpora are freely accessible over the World Wide Web. For example, casual users can search the British National Corpus online at <http://thetis.bl.uk/lookup.html>. There are even free Internet-based services which allow teachers and researchers to POS-tag their own English corpus texts automatically. Atwell et al (2000) reported that this opened up English corpus resources to a much wider audience, for example English language teachers set online classroom exercises on English grammar.

In addition, all the Arabic corpora available represent raw material or with very simple XML encoding except the General Scientific Arabic Corpus of which 1M words have been tagged with Parts of speech. There is another part-of-speech-tagged corpus which consists of 50,000 words and is based on newspaper texts (Khoja 2002). See appendix 1 for a sample of the corpus. However, this corpus is new and not (yet) in the public domain; furthermore, the size of this corpus is not large enough for some research purposes. In order to achieve a reliable result in many linguistic studies, the investigation has to be based upon a large corpus, which can be considered as balanced and as representative as possible of the linguistic community. Therefore, the aim of this project is to develop a corpus of Contemporary Arabic that would be of use for TAFL practitioners and would be freely available on the Web.

The next chapter describes a methodology for designing this corpus and examines the demand for such a project.

Chapter 3

Methodology

This chapter will describe the method used for planning the Corpus of Contemporary Arabic (CCA). Since the main aim is to produce a corpus for teaching Arabic as Foreign Language (TAFL), it was decided that the first step was to seek advice and opinions of teachers and language engineers on the best structure for the corpus so that the choice of text are not chosen randomly but rather to reflect the needs of the users. Sections 3.1 and 3.2 report on the methods used and the results obtained. Section 3.3 gives a brief description of the different forms of Arabic, and justifies the structure of the CCA.

3.1 Method Used

In March 2003 a survey of language teachers and language engineers was carried out to get their opinions on the texts that might be of use to them. An online questionnaire was developed and made available via mailing lists for language teachers and language engineers²⁹. It was also sent to some individual teachers. The questionnaire consisted of three sections. Section 1 contained personal detail questions covering the name of their company, nature of their business, name and contact address. Section 2 contained a list of 41 text types or genres which they were asked to rate on a scale of usefulness (very useful, useful, not useful). These texts belong to the following major categories:

Written: Fiction, Arts, Science, Business, and Miscellaneous

Spoken: TV, Radio and Conversation³⁰

Section 3 contained one question for language engineers and 14 questions for language teachers. The purpose of these questions was to examine the factors (if there were any), which affected their choice of texts, and to get their views on any other text that could be added (see appendix III for the complete questionnaire).

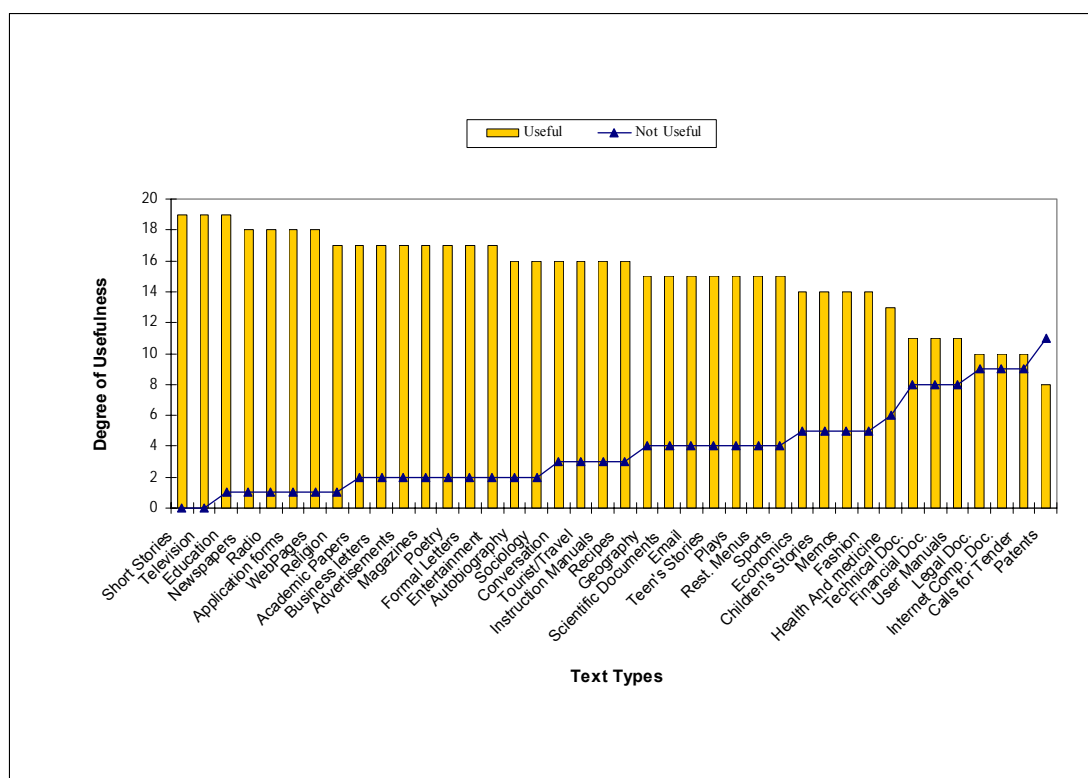
²⁹ The mailing lists used were: arabic-1@byu.edu , corpora@hd.uib.no , and elsnet-arabic@elsnet.org.

³⁰ The choice of text types was based partially on a survey conducted at the University of Leeds to find out the most frequently translated text types for the purpose of compiling a corpus for machine translation evaluation (Elliot et al 2003). More types were added from knowledge of other corpora.

3.2 Results and Discussion

30 replies had been received: 19 from language teachers, 11 from language engineers. The respondents were divided into the two groups and then a quantitative analysis was conducted. For the purpose of the descriptive analysis the ratings ‘very useful’ and ‘useful’ were grouped together to yield agreement frequencies. Both scores were positive and thus signal the importance of the texts for the corpus. Therefore two values had to be calculated: ‘useful’ against ‘not useful’. The responses from the language teachers were calculated to show their most useful texts. The same was done for language engineers. Figure 1 shows the scale of the useful texts, starting from the most useful to the least useful according to the language teachers’ opinions.

Figure 1: Distribution of the useful texts by language teachers



The graph shows that there is an overall consensus over the items ‘short stories’ and ‘television’: none of the language teachers rated these ‘not useful’. The remaining useful texts can be divided into categories based on their usefulness from the point of view of language teachers:

Category 1: short stories, TV, education, newspapers, radio, application forms, religion and web pages.

Category 2: academic papers, business letters, advertisement, magazines, poetry, formal letters, entertainments, autobiography, and sociology.

Category 3: conversation, tourist/travel, instruction manuals, and recipes.

Category 4: geography, scientific documents, e-mail, teen's stories, plays, restaurant menus, and sports.

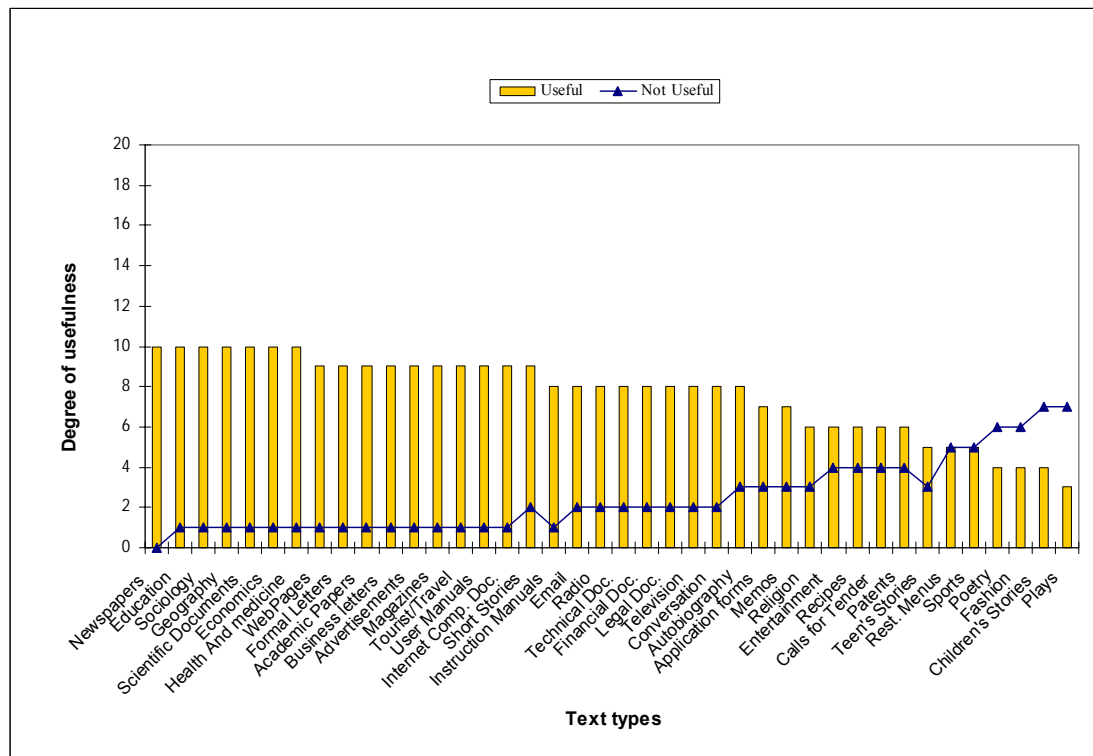
Category 5: economics, children's stories, memos, fashion, and health and medicine.

Category 6: technical documents, financial documents, user manuals, legal documents, Internet computer documents, calls for tender, and the text-type which is the least useful: 'patent'.

The result for language engineers shows that the most useful text for them is newspapers.

None of the language engineers rated this 'not useful'. Figure 2 shows the detailed result.

Figure 2: Distribution of the useful texts by language engineers



The rest of the texts can be divided into categories according to their classification by language engineers and their value of having equal usefulness. We should point out here that this classification of texts into categories is only made for ease of comparison.

Category 1: newspapers, education, sociology, geography, scientific documents, economics, health and medicine.

Category 2: webpages, formal letters, academic papers, business letters, advertisements, magazines, tourist/travel, user manual, Internet computer documents, short stories.

Category 3: instruction manuals, email, radio, technical documents, financial documents, legal documents, television, conversation.

Category 4: autobiography application forms, memos and religion, and patents.

Category 5: entertainment, recipes, calls for tender, and patents

Category 6: teen's stories, restaurant menus, sports, poetry, fashion, children's stories and plays.

Figure 2 highlights the expected pattern in that scientific and technical documents should be in the top categories. In the table they are in categories 1, 2 and 3 for language engineers, while for language teachers they came in categories 4, 5 and 6. But it was surprising that academic subjects were classified at the top of the list.

From the results it was possible to make the selection of the texts that should occupy the major part of the corpus. The texts that have been marked as less useful in both groups will be included but with fewer words. Even the less useful categories were judged "useful" by some of the respondents, so these should not be excluded entirely. Overall, the survey confirms that existing corpora are too narrowly limited in genre, and that there is a need for the CCA covering a broad range of text-types.

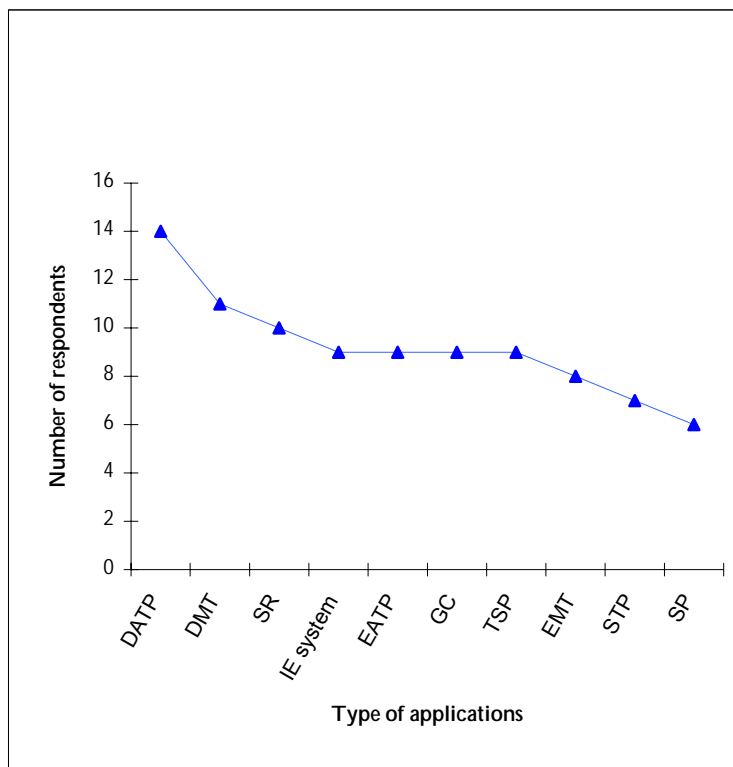
The questionnaire also asked the users to identify the potential future applications of the corpus and give their own suggestions for any other applications.

The ten potential applications suggested in the questionnaire were³¹:

- Developing Machine Translation (DMT)
- Evaluating Machine Translation (EMT)
- Information Extraction systems (IE)
- Developing Arabic text processing systems (DATP)
- Evaluating Arabic text processing systems (EATP)
- Grammar checkers (GC)
- Speech recognition (SR)
- Speech production (SP)
- Text to speech processing (TSP)
- Speech to text processing (STP)

³¹ Another application which was not included was developing new dictionaries for Arabic. With the advancement in English corpora and the production of new dictionaries based on corpora such as the COBUILD dictionary, it became obvious that traditional Arabic dictionaries are in need of updating. For instance, existing dictionaries such as Hans Wehr lacks new words and contains a great number of words which are not frequently used in MSA (van Mol 2000).

Figure 3: Number of responses for future applications of the corpus



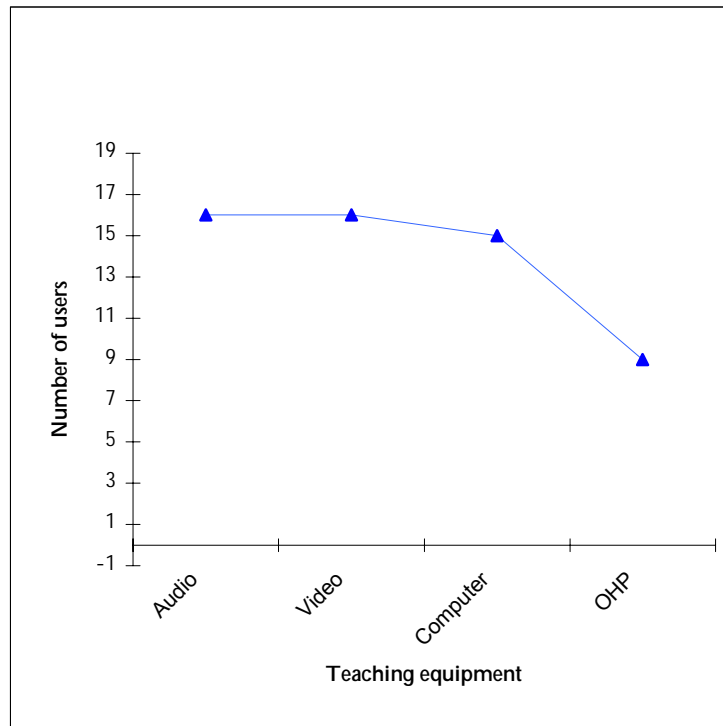
The applications that have been identified and for which the corpus could be a useful resource are shown in Figure 3. The second highest score of potential application for the corpus is ‘developing MT’. The graph in figure 3 shows 11 respondents out of 15³². This is a high score. This was interesting as it was planned that parallel Arabic/English texts should be included, but some justification or support from the users was needed. In addition to the support for Machine Translation applications, one respondent suggested using the corpus for translation studies. This purpose cannot be achieved unless some parallel texts were included. Furthermore, one question asked the participants to suggest other types of texts for the content of the corpus; and among the suggestions forwarded by the respondents were three suggestions by language engineers for including parallel texts. Based on the result in figure 3 and on the opinions of some of these respondents it was believed that including parallel texts in the corpus was as important as the other categories. Such texts are not only going to be useful for translation studies at advanced levels but also for studying grammar and learning about the distinctive structures of English and Arabic.

Another question was asked about the teaching equipment available for Arabic. The main purpose was to assess how much computers are used for teaching Arabic. If the result was

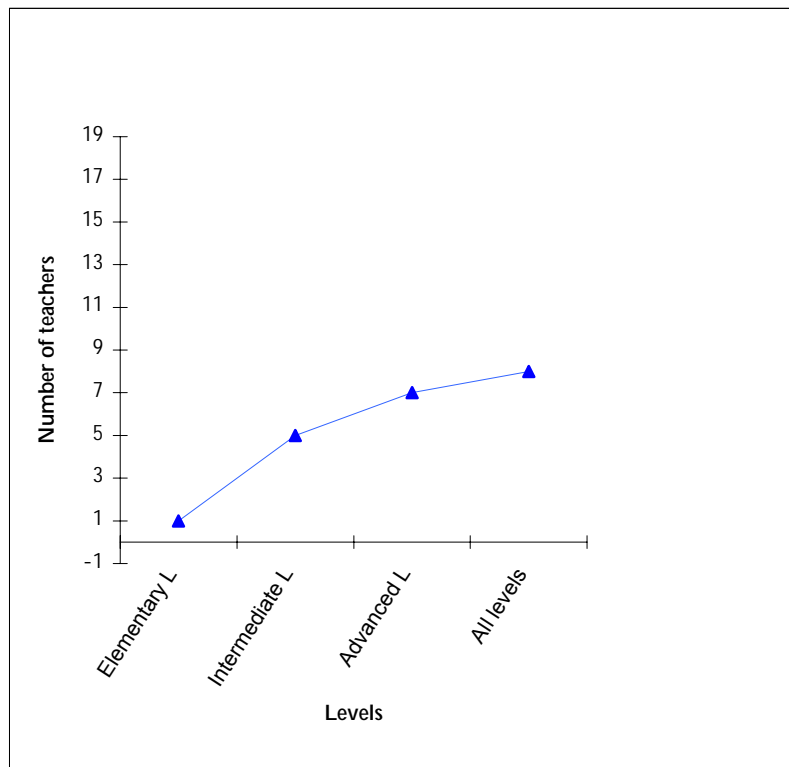
³² Only language engineers (of which there were 11) were instructed to answer this question but four language teachers answered it and so they were added.

high then there is potential in using the corpus as a teaching resource. Figure 4 shows, interestingly, that there is an increasing use of computers in the teaching of Arabic and less use of the OHP. The graph shows that there are 15 teachers out of 19 who use the computer in addition to other equipment.

Figure 4: Survey in the use of teaching equipment



One of the important issues regarding the content of the corpus was to include some written or spoken texts, which contain colloquial forms, as it is these types of text that represent contemporary Arabic. One possible source is Internet chat sites, which are characterised by their informality. It was not originally known whether such texts would be useful, so among the questions asked were whether teachers approve of teaching registers to foreigners. The results obtained from this question can be seen in figure 5.

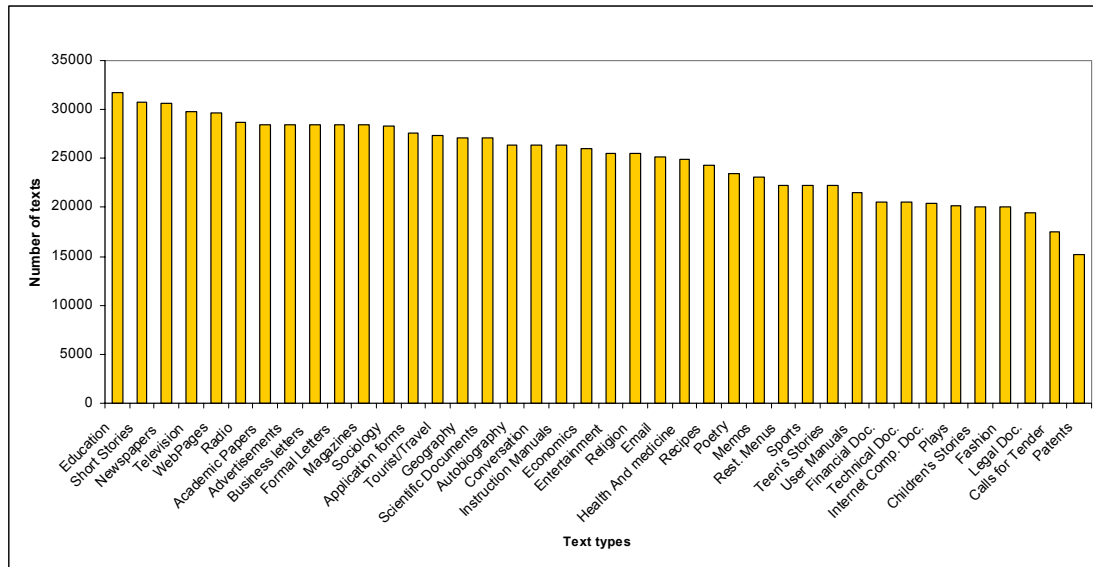
Figure 5: Number of teachers supporting teaching registers

Of 19 users, 17 agree that registers are useful for teaching foreigners. However, the highest score was for teaching it at advanced levels. At the same time the score for teaching it for all the levels was nearly as high as for using it for advanced levels. This signifies the importance of including colloquial forms in the corpus. In support of this finding, Lunt (1992) investigated the teaching methods used in five institutions in Tunis. She found that four of the institutions incorporate real data in their teaching either for reading or for listening. In her view, programmes that solely teach Modern Standard Arabic cause '*greater difficulty of application to the local environment*' (1992:122).

One limitation of this survey is the small number of replies that were received. It is well known, though, that people do not reply to questionnaires very readily. Brown (1988) points out that in research that depends on data collection '*...there is usually a certain amount of non-cooperation*' (1988:185). People do not cooperate fully especially to mailed questionnaires. Thus, such kind of method tends to yield a low response rate. It seems that online questionnaires, even though they reach a big number of people, have the same rate of response as mailed questionnaires. Some problems were encountered when checking the answers. One of these problems was finding some missing answers to some of the questions. In this instance answers were obtained by contacting the participants in person.

However, based on the survey the corpus was expected to contain the categories with the size shown in figure 6.

Figure 6: Number of words expected to collect for each category



This result was obtained by calculating the number responses of teachers and language engineers for every text type and then adding them up. Due to there being more responses from language teachers than engineers, the results had to be normalised first. However, since the main aim of the corpus was for the use of teaching, more weight was added for the teachers' answers by doubling their number. So for every text type there was a total number of responses. The total response for each category was then divided by the total number of the responses of all the categories, multiplied by the size of the corpus (1,000,000 words). This gives a percentage of the corpus size to be allocated to that text type.

Following the example of existing corpora (BNC) and due to the difficulty of processing spoken texts, a decision was made to have 10% of the CCA spoken and 90% written. The 10% spoken texts will cover as comprehensive a range of regional variants as possible. These variants of Arabic can be roughly divided into³³:

- Arabian Arabic which includes: a. Gulf Arabic (Qatar, Kuwait, Bahrain, Oman, United Arab Emirates, Bahrain, and Iraq), b. Hijazi and Najdi Arabic (Saudi Arabia), c. Yemen.
- North African Arabic: Libya, Algeria, Tunisia and Morocco.
- Levantine Arabic: Syria, Lebanon, Jordan, and Palestine.
- Nile Arabic: Egypt and Sudan.

³³ http://www.wordiq.com/definition/Varieties_of_Arabic

In addition to these groups, there are other speakers of Arabic elsewhere such as in Europe and other Islamic countries. But since Arabic is not the major language spoken in these countries, samples from such regions will be excluded (following the analogous principle in the International Corpus of English project, which only includes countries where English is the main language spoken). The ideal situation is to obtain equal proportions from each variety. That is, 25% from each, and also equal samples from each dialect within the regional variety. However, achieving this will depend mainly on finding some collaborators in that country which represent the regional Arabic; so far some researchers in Kuwait and United Arab Emirates have shown interest in this project and offered some help. A similar regional distribution of sources will be sought with respect to the written part. Although MSA is used in all these regions, there are slight variations in cases where the spoken is used within the standard. In order to study the linguistic differences between these regions in writing and to compare them with speaking, the written part must contain equal size. Thus of the 90%, there should be samples of 25% for each region. But this also depends on getting copyright permission from sources; and in practice it may prove difficult to pin down the regional origin of some sources, such as contributors to international newspapers.

3.3 Forms of Arabic

Arabic has three different forms: (i) Classical Arabic, which is the language of the Qur'an and classical literature; (ii) Modern Standard Arabic (or al-fusha), which is the language of newspapers and modern literature; and (iii) colloquial Arabic (or al-'ammiyya) which is the form of Arabic used in everyday oral communication. However, there is another form of Arabic referred to in linguistics by the term 'Educated Spoken Arabic, (ESA), 'al-lugha al-wusta' or the hybrid form. The characteristic of this form of Arabic is that it derives its features from the standard and the colloquial. Generally, it is used by educated speakers and also by speakers from one region when communicating with others from different regions.

Traditionally, it was believed that MSA is the ideal form to be taught for foreign learners because it is used in the Arabic media, and is also the common base for all Arabic dialects (Ferguson 1965). However, for the past twenty years or so spoken Arabic was regarded as important as MSA especially that the latter does not provide means of every day communication. Thus, there has been a debate over which dialect should be taught and should it be taught before MSA or after it. There are some who support teaching MSA before the 'ammiyya, while others support teaching both MSA and the 'ammiyya at the same time (Younes 1990). Still others support the teaching of ESA before MSA (Nicola 1990), or ESA after MSA (Haddad 1985). There is also variation in the regional or national varieties to focus

on; for example, a survey conducted by Elkhafaifi (2001) discovered that the most common dialect taught is Egyptian: 71% of instructors who answered his questionnaire teach Egyptian and the rest teach Moroccan, Syrian, and Palestinian. All these solutions have their advantages and disadvantages. However, the problem with the ESA, which the other forms do not have, is that its form is not yet defined. It varies from one region to another. It might even vary from one person to another. Despite that, we cannot deny its existence and the fact that it is used in our daily communication.

Holes (1990) pointed out how the teaching of Arabic to foreigners does not seem to reflect the reality of the language. There is a great emphasis on teaching students how to read and write and translate or criticise pieces of classical literature but there is no opportunity for students to be exposed to the contemporary reality of the Arabic language. As he states:

'...the reality, for example, that while people write fusha they may speak with a variety of regional and social accents, the reality that while they may read or listen to an expose about a subject in fusha or colloquial, they will talk about it in the latter and write about it in the former' (1990:37). He suggested that the emphasis should be *'...on using authentic material from a variety of contemporary sources for authentic ('real life'-like) purposes'* (ibid).

The rationale of the CCA is based on this stance. MSA is not the only form foreigners should be exposed to. They need to be exposed to contemporary and 'real' Arabic in addition to MSA. This Arabic is represented in political speeches, plays, interviews, emails, Internet discussions, chat sites, etc.

Therefore, the term 'Contemporary Arabic' can be defined by the form of MSA used across the Arab speaking countries which is written or spoken in the 1990's up to the present time, as well as contemporary regional varieties and ESA. In teaching foreign learners it is practical to choose one specific variety along with the MSA. Students would know some of the basics of the dialect or even more by spending one year in an Arabic country as is the case in some Arabic degree programs such as the one in the University of Leeds. However, although native speakers would understand him/her but it does not guarantee that the learner would understand people who speak other varieties. Even though Egyptian dialect is traditionally known to be the most understood dialect by the other speakers, this does not mean that Egyptians themselves understand speakers of other varieties, especially if they did not have contact with people outside Egypt. In most cases speakers of other varieties try to speak Egyptian so that they can be understood. This means that students who learn a particular dialect expect every Arabic speaker to speak that dialect if he or she would communicate

effectively. Highly educated people are equipped with knowledge to modify their speech to ease communication with foreign speakers but common people are not. Therefore, as Thomson (1994), being a teacher of English to Arab-speaking students and a learner of Arabic himself, suggests:

'to remedy this situation, the aim of the colloquial Arabic course in the final year(s) would be to introduce the students to a broad range of different varieties of spoken Arabic. The emphasis of the course would be on listening comprehension; the goal would be recognition of different forms rather than their production. The students would thus gain a better appreciation of common dialectal features and of which features of the dialect they have studied are peculiar to it and which have more general applicability' (1994:18).

Nowadays, with the appearance of the satellite TV, people have access to all the different channels and thus different spoken varieties. Recently, it has been reported that the Lebanese channels LBCI (International Lebanese Broadcasting Corporation) and Future Television have a wider audience than other channels. This means that the Egyptian dialect is no longer the dominating variety.

Therefore, the plan of the CCA was to reflect the reality of the Arabic language in order to help learners in foreign countries to have a wider view of how Arabic is used. The main focus would be representing the MSA form written and spoken as well as some regional varieties which can be reflected in radio and TV programmes. It was hoped that the corpus would be a rich resource for learners to explore, compare and learn about the present Modern Standard Arabic with its new vocabulary and its different regional varieties.

3.4 Teaching Arabic in the 1990's and the Future

The incorporation of new technology in teaching in general has its big effect on the teaching of foreign languages. Since the nineties there has been some effort in introducing computers, videos and multimedia material in the teaching of Arabic. Although it is done on an individual basis and still limited, there is an awareness of the importance of making use of students' technological skills in their learning of Arabic. Methods and approaches of teaching have improved and traditional methods such as grammar-translation have been abandoned. Learners of Arabic are no longer taught solely the standard as it used to be but they are taught the dialect as well. This is achieved by learning the dialect of the Arab country they are studying in or by traveling to an Arab country to study and practice the dialect. The current Arabic programmes in the UK and US do send their students abroad to learn the dialect after

spending a year or two studying the standard. This is a good opportunity for them to absorb the culture and learn the spoken variety in its natural environment. This change in the philosophy of teaching Arabic is related to the awareness that the main purpose of learning a language is to communicate orally as well as to read and write.

In a national survey conducted by Belnap (1987) of five hundred students learning Arabic at college level in the US, he found out that the learners' priority in learning Arabic is to communicate with Arabs and learn about the Arabic culture. In this survey he also asked students, as well as teachers, about their attitudes in learning a dialect. The result has shown that 30% of students thought it is desirable to learn the dialect in comparison to 6.7% who thought it is undesirable. The teachers also agreed except that they disagreed on the timing of introducing it. One of the interesting findings in this survey is that in a two-year programme for teaching Arabic it was found that only less than 20% of students continue beyond the second year and 80% of students do not have a chance to learn the spoken language. This implies that learners are frustrated because they are not using what they learn for communication. Therefore, such findings put pressure on institutions to change their Arabic programmes to match the learners' needs.

There is at present a new trend of adopting a communicative approach over the traditional one and of incorporation of the cultural aspect in the Arabic programme. Elgibali and Taha (1995) describe the programme of teaching Arabic as a foreign language in the Arabic Language Institute (ALI) at the American University in Cairo, as being more consistent with the way Arabic is used. There is an emphasis on using authentic material in both reading and listening. Topics cover current issues in politics, economics, history and culture. In listening, learners are exposed to recordings of both fusha and colloquial. This is to prepare them to understand people outside the classroom where native speakers combine the two to use for speaking. This approach is used for both beginners and advanced. They differ only in the degree of difficulty and length of dialogues. As for speaking, although it is hard to choose which dialect to teach, they stress the teaching of a dialect for this skill and it is preferable to teach the dialect of the country they are living in. This is followed by teaching how to speak in al-fusha, like preparing them to deliver a public talk and being able to carry out a discussion. This approach is considered to be more realistic and fit in with the way Arabic is used by its native speakers. To teach learners to converse in the standard Arabic is not only considered unnatural but the majority of native speakers themselves are not skilled at using it.

Al Batal (1995) also discussed the situation of teaching Arabic in classrooms and of how to make it fit in with the real situation of Arabic as spoken by its native speakers. To him, Arabic

is not just standard or not just dialect. It is both integrated as one entity. He proposes the teaching of MSA as written and spoken as well as a dialect and exposing students to another variety in which the standard and the dialect are integrated. He points out that this is a big step towards improving the quality of teaching Arabic and responding to students' needs. But this also requires more research into the similarities and differences between the standard and the dialects to help in the development of new textbooks in the way that make learning the two forms of Arabic much easier. As a reaction to the proposal of integrating the standard and the dialect in teaching, several textbooks have already been published and in which an introduction of a dialect had been included with the standard. For example, Munther Young produced an elementary book in which MSA is introduced with Levantine Arabic. Brustad, Al-Batal, and Al-Tonsi produced a book for teaching MSA but with some reference to Egyptian colloquial Arabic. Al-Batal as well stresses the importance of teaching writing as it is another skill for communication. Even though 44% of students surveyed by Belnap rated writing as the least important learning skills, it needs to be given more attention, and there is an urgent need for producing textbooks to teach foreigners how to write effectively. At present there is more focus on teaching grammatical sentences and avoiding errors but focus on the mechanics of writing and discourse levels is ignored.

Taha (1995) questions the grammatical rules presented in textbooks. She believes that these rules reflect our views of how MSA should be used but they do not reflect the real rules as used in media, formal speeches, and literary works. There are lots of variations in the media which are not described or accounted for. She gives evidence of the different use of conditional clause in the Egyptian newspaper *Al-Ahram* and the Saudi newspaper *Al-Sharq Al-Awsat*. All MSA books state that the rule of conditional clauses with /IDa/ and /law/ should have /fa/ and /la/ in their response clause. This statement is not completely accurate as it is possible not to have /fa/ and /la/ in the response and this is evident in the above newspapers. She expresses her view by saying that:

'We regard MSA as what we want it to be rather than what it is in reality. More importantly, I think we cannot even realize that what we see today may not be the same as what we will see in the future and it is definitely different from what we used to see in the past (1995:180).'

She warns that if we continue to give students rules of MSA as derived from Classical Arabic then we are not presenting the rules of MSA as it is used today and not accounting for variations and forcing students to learn rules not used by the native speakers themselves.

Alosh (1995) and Parkinson (1995) explore the potential use of the computer in teaching Arabic and discuss the design and development of CALL (Computer Assisted Language Learning). They stress that programs should not only suit different levels and interests of learners but also have a broad applicability so that they can be used in different Arabic programs because producing such programs demands lots of work and resources.

Hoogland (2003), in his experience of developing an Arabic-Dutch dictionary observed that editors who have been hired as native speakers to work on the dictionary tended to consult each other, dictionaries or the corpus for more confirmation. In his view this shows that there is no genuine native speaker of MSA and this is due to the fact that native speakers of Arabic learn a dialect before they start learning the standard and some Arab speakers could even have a non-Arabic language like Berber or Koptic. Also, since MSA is not the language used in daily life it was hard to find translation in Standard Arabic for lots of expressions which are used in spoken language. It was interesting to know that this dictionary contains 'a considerable number of expressions which could best be classified as 'spoken language' and these are translated by detailed description as there is no equivalent for them in MSA.

This new trend in teaching is analogous to current thinking in English language teaching. British English TEFL as promoted by the British Council used to focus more on RP (Received Pronunciation) and 'standard' southern British English or BBC English (although it was never really an officially recognised standard in the same way as standard educated Arabic, or French). Yet, a recent survey of TEFL practitioners (teachers, publishers, students) showed that many advocate exposure to a range of British English dialects, since a learner visiting Britain would have to contend with regional dialects in reality (Atwell et al 2000).

However, which form of Arabic to teach is still a contentious issue. Some 'traditionally-oriented' Arabic language teachers still prefer to teach only MSA. Despite that, there is a need for a corpus of contemporary Arabic (CCA) to cater for the growing alternative view.

This project will be another new resource to add to the new textbooks which had been published recently and which would help in the progress and improvement of the quality of teaching resources.

Chapter 4

Corpus Encoding

Whether the corpus is written or spoken two types are identified: raw corpus and annotated (or marked-up) corpus. The former is mainly the text itself with no other additional information and the latter the text is enriched with a variety of information. The purpose of a corpus is to use it with the help of special software to investigate the structure of the language and to extract other types of information for the purpose of teaching and research. Although raw corpora can be used with the help of tools to investigate any kind of linguistic analysis, in order to be able to get more refined information such as whether the linguistic features are related to the social class, age, or sex, the corpus must be encoded using some sort of mark-up language to enable the user to extract this information. The information that can be encoded includes linguistic and non-linguistic features. They are such as:

- Paragraphs, sections, headings, sentences.
- Boundary of part of speech of each word.
- Speech turns, pausing.
- Paralinguistic features such as laughter and hesitation.
- Meta-textual information such as the source of the text, author, publishing company, etc.³⁴

The following is a sample of a piece of BNC text with part-of-speech markers and other types of boundaries (taken from 'ICI Chemicals Polymers: Environmental issues')

```
<p>
<s n="3"><w CJS>When <w PNP>it <w VVZ>comes <w PRP>to <w VVG>cleaning
<w AVP-PRP>up <w AT0>the <w NN1>environment<c PUN>, <w PNP>it <w
VBZ>is <w VVN>said <w CJT>that <c PUQ>&#34;<w AT0>the <w
NN1>polluter <w VVZ>pays<c PUQ>&#34;<c PUN>&#34;<w CJC>but <w
PNQ>who <w VBZ>is <w AT0>the <w NN1>polluter<c PUN>?
</p>
<p>
<s n="4"><w VM0>Let's <w XX0>not <w VVI>fool <w PNX>ourselves<c PUN>.
```

³⁴ More information is at <http://www.natcorp.ox.ac.uk/what/encoding.html>


```
<s n="5"><w CJS>When <w NN1>electricity <w NN2>generators <w VVB-
NN1>tackle <w NN1>acid <w NN1>rain <w PRP>with <w
NN1>desulphurisation <w NN1>equipment <w CJC>or <w NN1>water <w
NN2>authorities <w VVB>clean <w AVP>up <w NN2>beaches <w PRP>by <w
AJ0-VVG>improving <w NN1>sewage <w NN2>treatments <w NN2>plants<c
PUN>, <w PNP>we <w DT0>all <w VVB>end <w AVP>up <w VVG>paying <w
PRP>via <w DPS>our <w AJ0>domestic <w NN2>bills<c PUN>.
</p>
```

4.1 XML and Corpus Encoding

There are several mark-up languages that can be used but the most efficient is the one that creates a compatible encoded document. This is crucial so that the encoded corpus can be used and processed regardless of the different computer systems without losing information as a result of file transfer. In 1986 the Standard Generalized Markup Language (SGML) was considered to be an international standard for defining different types of electronic documents and it was popular in many organisations. However, although it is a flexible language, there are several problems that prevented it from transferring text over the Web. Some of these are lack of general stylesheets and problems of interchanging data because of using different software packages. HTML which is an application of SGML was not the best alternative as it contains predefined and restricted tagsets which makes difficult to process documents with complex structure. It also creates loss of flexibility and prevents the automatic interchange of files. For all these factors XML (Extensible Markup Language) was developed by the XML group known as SGML Editorial Review Board which is formed under the auspices of the World Wide Web Consortium (W3C) in 1996. XML is a set of conventions to use SGML without its complex features.

XML is concerned with two types of components:

- a. low level component that contains features such as headings, paragraphs, sentences and type of document whether it is a book or a report.
- b. high level component that contains features that are used for encoding the document. These features are known by Document Type Definitions (DTD). It defines the structure of the document and its elements. For example, there are different elements for marking up documents depending on its type whether it is poetry, or novel or drama.

Corpus builders should stick to the guidelines of DTD. It is used by XML parser to analyse a document automatically and check that it complies with specific DTD. This process is called validation. For example, if a document type definition for a book specifies that it has a title, author, table of contents, an x number of chapters and an index, only documents meeting these features can be grouped as books. DTD is also used for checking structural errors in the document. DTD is very complicated and it needs a certain amount of time to create it. There

are also other guidelines in DTD that gives a standard format for documenting all the information necessary about a text. This can be referred to as the Text Encoding Initiative (TEI) and it is represented as a header attached at the top of each document. To conform to the guidelines of TEI, each document in a corpus must include a header that contains some basic information about the title of the text, the author, the publisher, and much other information. For example, the texts encoded in the BNC contain a great amount of information. Recently there was a discussion among the Corpora Mailing List's members about the value of TEI and whether it is a waste of time. The general consensus is that it is an essential part of designing a corpus. The more accurate and detailed the markup is the easier it is to use the corpus for generating the information users want. A corpus is not merely a compiled selection of texts. Rather the texts need to be organised and stored with their general and specific features so that when a linguistic query is conducted the correct information is returned. The organisation of this information is achieved by TEI.

The problem facing corpus developers is lack of sufficient software for encoding the markup. There is no public framework or set of tools for treatment of TEI-corpora. The present situation is that some develop their own simple and quick TEI-XML specific tools while others decide on writing minimal encoding and pasting a header using a word processor. For this project the latter approach has been adopted, manually encoding what was necessary for the purpose of the corpus. Therefore, the bibliographic information and socio-linguistic parameters associated with the text have been included.

Generally, there are four components that are necessary for each document. These are:

File Description <fileDesc>

It is an obligatory element in the header, and it includes bibliographic information about the text such as title of the work, name of the author and publishing company. Below is an example:

```
<teiHeader>
<fileDesc>
<titleStmt><title>Data Mining: practical machine learning
tools and techniques with Java implementations</title>
<author>Ian H. Witten, Eibe Frank</author>
</titleStmt>
<publicationStmt>
    <publisher>Morgan Kaufmann</publisher>
    <pubPlace>San Francisco</pubPlace>
    <date>2000</date>
</publicationStmt>
<sourceDesc>
```

```
<bibl> Data Mining: practical machine learning tools and
techniques with Java implementations by Ian H. Witten and Eibe
Frank (San Francisco, 2000) </bibl>
</sourceDesc>
</teiHeader>
```

The above is the minimal description of a text. Due to the limited time available for this project this will be the adopted format of the CCA.

Encoding Description <encodingDesc>

It states the relationship between the text and its source, and it contains nine optional subdivisions. For the CCA two elements were chosen from this category. These are the project description and the sample declaration. It is essential to state the aim of the project and some detailed information about the sample.

Below is an example of how they are stated (Baker et al 2003):

```
<projectDesc>Text collected for use in EMILLE
project</projectDesc>
<sampleDesc>simple written text only has been transcribed.
Diagrams, pictures and tables have been omitted and their
place marked with a gap element</sampleDesc>
```

Profile Description <profileDesc>

It supplies non-bibliographic information about the text and the participants. It contains the following elements: creation, language usage, text class, text description, participant description, and setting description. Since they are optional it was decided only those deemed relevant to the corpus would be included. The elements that have been chosen are:

1. Text description <textDesc>: this element provides information regarding the medium by which it is delivered. That is, whether it is print, email, face-to-face, TV, etc. Also whether it is written or spoken, spoken to be written, written to be spoken. Its derivation must be stated whether it is original or translated. It is important to state the domain such as art, religion, history, etc. In general, all what is available about the text must be included. However, in the situation where information could not be found, the entry would be 'unknown', or if it does not apply, 'inapplicable'. Below is a sample³⁵:

```
<textDesc n='novel'>
```

³⁵ It is derived from <http://www.hti.umich.edu/cgi/t/tei/tei-idx?type=HTML&rgn=DIV3&byte=1984593>.

```

<channel mode=w>print; part issues</channel>
<constitution type=single>
<derivation type=original>
<domain type=art>
<factuality type=fiction>
<interaction type=none>
<preparedness type=prepared>
<purpose type=entertain degree=high>
<purpose type=inform degree=medium>
</textDesc>

```

2. Participant description <participantDesc>: information about participants in the text is important. Participants can be an author, speakers in a dialogue, an interviewer and interviewee, etc. The information that needs to be provided consists of sex, age, nationality, date of birth, place of residence, languages spoken, education and occupation. In addition, it is necessary that each participant takes an identifier to make it easy when processing the texts and searching for certain people who belong, for example, to a certain social culture. A sample is given below³⁶:

```

<person id=P1 sex=F age='mid'>
  <birth date='1950-01-12'>
    <date>12 Jan 1950</date>
    <name type=place>Shropshire, UK</name>
  </birth>
  <firstLang>English</firstLang>
  <langKnown>French</langKnown>
  <residence>Long term resident of Hull</residence>
  <education>University postgraduate</education>
  <occupation>Unknown</occupation>
  <socecstatus source=PEP code=B2>
</person>

```

Revision description <RevisionDesc>

It gives a summary of the history of the text or the book. Thus, if it is updated at some point, there should be a mention of the date and the state of its modification. It is an important element to add to the corpus. However, it is not easy to go through the history of each document obtained.

³⁶ It is derived from <http://www.hti.umich.edu/cgi/t/tei/tei-idx?type=HTML&rgn=DIV3&byte=2001380>.

4.1.1 Summary

For the CCA, most of this information was included depending on how easy it was to access it. For Arabic resources, sometimes even the essential informational of dates and publishing sources are not readily available on the Web.

In addition to the low level encoding of the corpus, there are other types of high level annotation, such as part-of-speech, syntactic, semantic and prosodic annotation. The important one is part-of-speech annotation. To make the corpus more useful for teaching and research it is essential to assign words their grammatical categories so that also it would increase data retrieval and disambiguate homographs. In Arabic, since vowels are not connected with words, it is highly difficult to use especially for foreign learners. Thus, having the grammatical tagging would simplify word recognition and identifying their meaning. Unfortunately, at this stage of the project this aim can not be pursued. The encoding will be very basic and limited to only assigning paragraph markers and the header.

See appendix IV for a sample of the header adopted for the corpus.

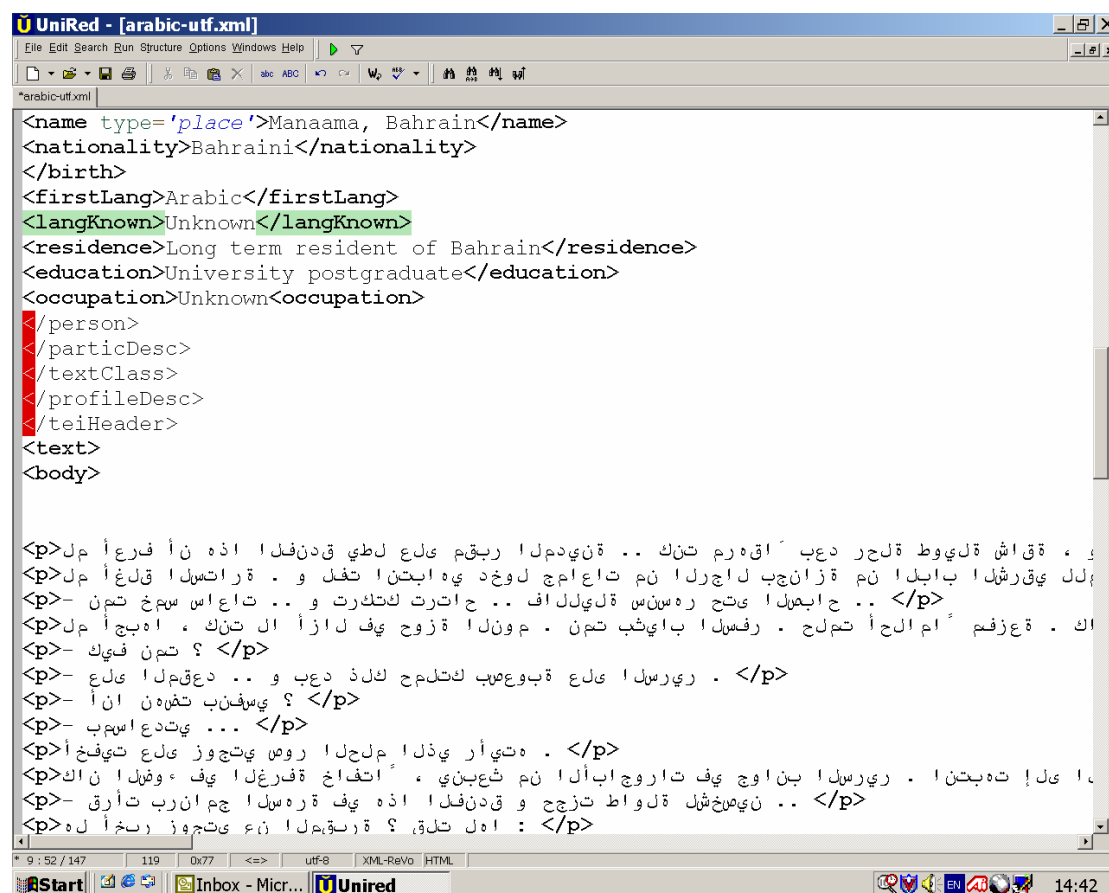
4.2 Procedure of the Corpus Encoding

Annotating large corpora is usually done by means of some computer programs so that a many texts can be annotated in a short time. It is normally done automatically and then if there are any mistakes they are corrected manually. Unfortunately, there was no means of developing a program to encode this corpus automatically, therefore, encoding was performed manually. Below is a sketch of the overall procedure:

1. Copy a text from the Internet and paste it into Microsoft Word. Then save it as encoded text choosing Unicode UTF-8.
2. After counting the words in the text, encode the text with paragraph marker using the option FIND/REPLACE in edit: Find ^p Replace </p>^p<p> in the case of a normal text and Find ^p Replace </l>^p<l> in the case of a verse as it contains lines.
3. After making a decision of the elements needed to include in the corpus, a template of the header was created and saved.
4. After the paragraphing was marked, the text was prefixed with the template of the header and suffixed with the closing brackets. A name was given based on the text category. Thus, a story could be named S01.txt.
5. Fill in the specific information about the text and the author.
6. Rename the text by changing the file extension from .txt to .xml.
7. Open the text in Internet Explorer to verify the XML file.

In the event that there is a problem with viewing the file for one reason or another, it was opened again in a program called UniRed (it is a Unicode plain text editor for windows and it supports many character sets including UTF-8 and mark-up languages such as XML and HTML). This program identifies the errors by highlighting them with red. If an XML tag appears green it is correct and if it is red it means it is invalid. Sometimes the mistakes are related to the coding of the header but this is very rare since the template is checked ahead of time. Other times the mistakes are related to some unusual characters or signs which need to be modified to be accepted by XML. (e.g. & sign which needs to be written as 'and').

Figure 7: Screenshot of a sample from UniRed editor



```

UniRed - [arabic-utf.xml]
File Edit Search Run Structure Options Windows Help
*arabic-utf.xml
<name type='place'>Manaama, Bahrain</name>
<nationality>Bahraini</nationality>
</birth>
<firstLang>Arabic</firstLang>
<langKnown>Unknown</langKnown>
<residence>Long term resident of Bahrain</residence>
<education>University postgraduate</education>
<occupation>Unknown</occupation>
</person>
</particDesc>
</textClass>
</profileDesc>
</teiHeader>
<text>
<body>
<p> .. عايشة ليوطة لحر دعب اقهرم تنك .. عني دملا ربقم لعل لطي قدنفلدا اذه نأ فرع أ مل</p>
<p> لعل يقرشلدا بابلا نم ةزانجب لاجرلا نم تاعامج لوخد يه ابتنل تغل و .. ةراتسلدا قلغ أ مل</p>
<p> .. ح ابعلدا ايتح ره سنس ةليللاف .. ح اترت كتكرت و .. تاعامس سمخ تمن -</p>
<p> اك .. ةعزفم املح أ ملح .. رفسلدا بابي شب تمن .. مونلا ةزوح يف لازا امل تنك ، اه بجا مل</p>
<p> -</p>
<p> .. ريرسلدا لعل ةبوعصب كتلمح كلذ دعب و .. دعقملدا لعل -</p>
<p> .. يسفنن ب تهنه ان ا -</p>
<p> .. يتدع اسمب -</p>
<p> .. هتي أري ذل املحلا روص يتجوز لعل تي فخ أ</p>
<p> .. ريرسلدا بناوچ يف تاروچ ابألا نم شعبيني ، اتفاخ ةفرغلدا يف ءوضلا ناك</p>
<p> .. نيصخشل ةلواط تزجج و قدنفلدا اذه يف ةره سل اجم انرب تارق -</p>
<p> : اهل تلوقة ؟ ةرسملا نء يتجوز رسخأ له</p>

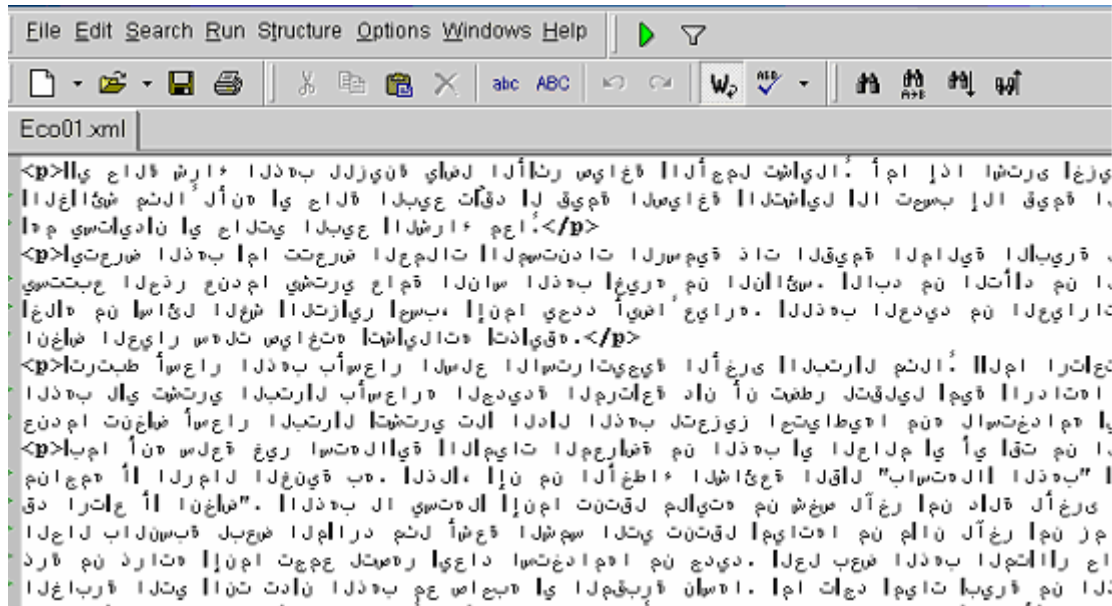
```

4.3 Some Specific Problems

This section will discuss some specific problems which were encountered in the process of collecting texts from the WWW. Firstly, it should be noted that the method described in the previous section gave valid results for the majority of texts. However, on occasions, there were problems that cause the XML file to be invalid.

A couple of problems were discovered when opening the text in UniRed. The first one was that sometimes Arabic text appears with vertical bold strokes in between the letters. These strokes prevent the text from being viewed in the browser. The figure 8 demonstrates the problem.

Figure 8: Screenshot showing a distorted file



When this text was copied and pasted into a word processor, it showed that the vertical bold line is some kind of distorted letter with four dots which does not exist in the language. The text below gives a demonstration:

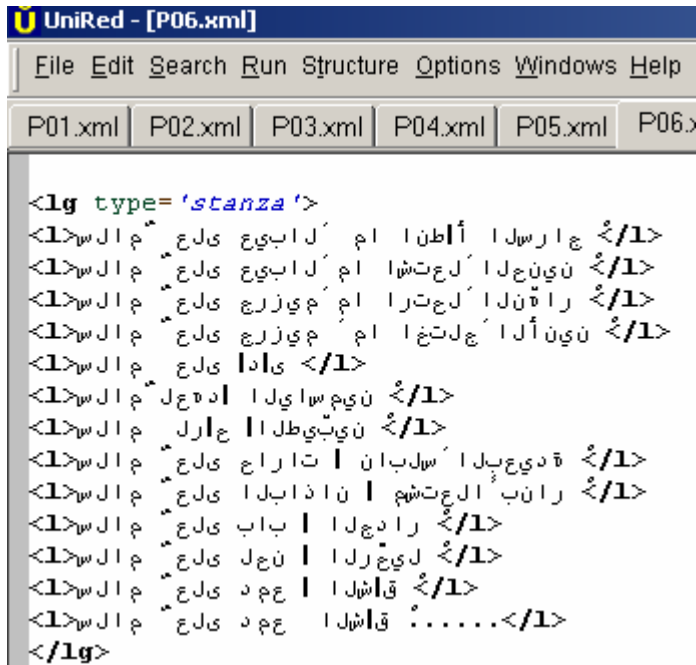
تتي حالة شراء الذهب للزينة يتضل الأتتر
 صياغة تالاجمل تشتيلا. أما إذا اشترى تخزينة،
 أي للادخار، تيتضل شراءه تي شتل سيائت أت
 غيرها من أشثال الذهب القليل الصياغة
 تالغتائش مثلا لأنه تي حالة البيع تُتقد تل قيمة
 الصياغة تالتشثيل تلا تحسب لإقيمة الذهب فقط
 حسب تزنه، مما يجعل أرباح التجارة تيه أتتر
 ضماناً من التجارة تي غيره. تهم يستتيدتن تي
 حالتي البيع تالشراء معاً.

The text below represents the original:

وفي حالة شراء الذهب للزينة يفضل الأكثر
صياغة والأجمل تشكيلاً. أما إذا اشترى
كخزينة، أي للادخار، فيفضل شراءه في شكل
سبائك أو غيرها من أشكال الذهب القليل
الصياغة كالغوائش مثلاً لأنه في حالة البيع تُفقد
كل قيمة الصياغة والتشكيل ولا تحسب إلا قيمة
الذهب فقط حسب وزنه، مما يجعل أرباح
التجارة فيه أكثر ضماناً من التجارة في غيره.
فهم يستفيدون في حالتي البيع والشراء معاً.

This means that the letters ك و ف are the ones which are distorted. This problem occurs in texts from some sites among them the Economic World Magazine³⁷ and Ofouq magazine³⁸. There are some files similar to these especially in poems. In addition to the above letters, texts that have the vowels 'Harakaat' create problems. The kasra /i/ appears as a bold vertical line. The example below shows the difficulty in identifying it:

Figure 9: Screenshot showing a file with Harakaat 'vowels' distorted



```

<lg type='stanza'>
</1> ج ار سدا اأطن ا ام ل ا ب ي ج ي ل ع م ا ل س </1>
</1> ن ي ن ج ل ا ل ع ت ش ا ام ل ا ب ي ج ي ل ع م ا ل س </1>
</1> ر ا ة ن ل ا ل ع ت ر ا ام م ي ز ر ج ي ل ع م ا ل س </1>
</1> ن ي ن ا ل ا ل ع ل ت غ ا ام م ي ز ر ج ي ل ع م ا ل س </1>
</1> ي ا د ا ي ل ع م ا ل س </1>
</1> ن ي م س ا ي ل ا ا د ع ل م ا ل س </1>
</1> ن ي ت ي ط ل ا ا ع ا ر ل م ا ل س </1>
</1> ة د ي ع ب ل ا س ل ب ا ن ا ت ا ر ا ج ي ل ع م ا ل س </1>
</1> ر ا ن ب ا ل ع ت ش م ا ن ا ذ ا ب ل ا ي ل ع م ا ل س </1>
</1> ر ا د ج ل ا ا ب ا ب ي ل ع م ا ل س </1>
</1> ل ي ج ر ل ا ا ن ج ل ي ل ع م ا ل س </1>
</1> ق ا ش د ا ا ع م د ي ل ع م ا ل س </1>
</1> ق ا ش د ا ا ع م د ي ل ع م ا ل س </1>
</lg>

```

³⁷ Available at <http://www.ecoworld-mag.com/>.

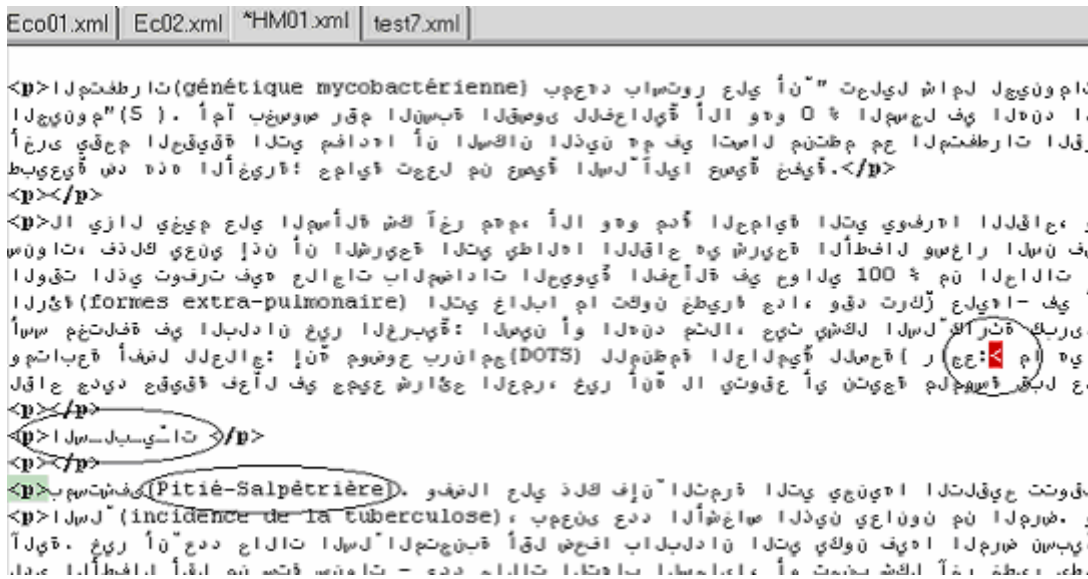
³⁸ Available at <http://www.ofouq.com>

The original text with the Harakaat properly displayed:

سلامٌ على عيالٍ ما انطفأ السراجُ
 سلامٌ على عيالٍ ما اشتعلَ الحنينُ
 سلامٌ على جرزيمٍ ما ارتحلَ النهارُ
 سلامٌ على جرزيمٍ ما اختلجَ الأنينُ
 سلامٌ على فدوى
 سلامٌ لعهدِ الياسمينِ
 سلامٌ لروحِ الطيبينِ
 سلامٌ على حاراتِ نابلسِ البعيدةِ
 سلامٌ على الباذانِ مشتعلاً بنارِ
 سلامٌ على بانهِ الجدارِ
 سلامٌ على لحنِ الرَّحيلِ
 سلامٌ على دمعِ الشفقِ
 سلامٌ على دمعِ الشفقِ.....

Another problem was that the file contains some symbols such as < or it contains some French words or some unique shape of Arabic font where there is a line between each letter to make the letter long. This is used sometimes for titles. UniRed identifies some of these problems such as the < but not others. These create problems as they cannot be removed by find and replace. They need to be deleted manually and so it is time consuming and it does not even guarantee that the file is clear of all these unwanted shapes. This problem was encountered in some sites such as the Arab Medical Magazine³⁹. Figure 10 exemplifies this.

Figure 10: Screenshot showing a file with unaccepted characters



³⁹ Available at <http://arabmedmag.com/>

These problems caused some frustration and slowed down progress. At one stage it was decided to start deleting all the corrupted files or those that can not be viewed in the browser. The reason for this corruption could not be detected. However, an alternative method was tried which created the template of the header in UniRed. The text was then inserted within the editor (rather than using a word processor) and then the header was updated to reflect the information about the text. This method proved not only faster but also did not create distorted files.

4.4 Processing Written Texts and Speech Recordings:

4.4.1 Written Texts:

During the text collection stage the time taken for processing was calculated. Using the method described previously, the time usually taken to go through the steps was approximately 15 minutes. If further files from the same site were collected, the header could be reused with some minor adjustments to fit the new text. This would obviously take less time than the first file. It ranges between 6 to 10 minutes depending on whether the text is void from any strange symbols. This is the time usually taken when creating the text in Microsoft Word. However, when creating the text in UniRed, it was not only faster but also guaranteed that the files were correctly saved and can be viewed in the browser with no problems. Processing time with this method was 3-5 minutes.

4.4.2 Spoken Recordings

It was believed that having a good amount of spoken recordings was fundamental in the design of the corpus. The plan was to contact some Arabic radio stations and obtain some spoken recordings covering topics such as general speeches, news, interviews, plays, narrations, poetry reciting, phone-ins and other programmes that might be useful. However, speeches such as preaching would be excluded as this kind of genre contains Classical Arabic. Responses were received in the form of CDs and audiocassettes. CDs were the preferred medium as the recordings were already in digital form. But if the recording is on audiocassettes, they had to be digitised first.

For the CCA the spoken recordings were transcribed orthographically. However, transcribing Arabic spoken recording is not very easy, especially when using Arabic script. It was the most laborious stage in the project. Processing many files in quick succession does not make the work faster as the exhaustion slows down the process and it is preferable to take a rest between each file. The first recording that was transcribed had the duration of 5 minutes, yet took 1 hour and 5 minutes. It was difficult to input it directly onto the computer. So the

recording was first handtranscribed - this took 30 minutes. Then it was typed up, which took 35 minutes. A professional Arabic typist would have been able to process the recording in a single step and would be much faster. Adding the header with the necessary information took 8 minutes. Table 4 gives a summary of all the recordings transcribed.

Table 4: Records of spoken texts and length of time of transcribing them

<i>Text type</i>	<i>File no.</i>	<i>Duration</i>	<i>No. of words</i>	<i>Time to transcribe</i>	<i>Speed per min</i>
Monologue	Edu01	5 mins	435	1 hour 5 mins	6 w
Monologue	Edu02	7 mins 26 sec	805	1 hour 47 mins	7 w
Interview	Spo01	4 mins 51sec	682	2 hours 30 mins	4 w
Interview	Spo02	3 mins 42 sec	535	2 hours 4 mins	4 w
Conversation (soap opera)	Entr02	8 mins 26 sec	1377	2 hours 43 mins	8 w
Interview	Spo06	3 mins 38 sec	519	2 hours 2 mins	4 w
News	Pol02	12 mins 31 sec	1252	2 hours 15 mins	9 w

As table 4 illustrates, the duration of the time does not depend on the length of the text but rather on its type. It is easier to recognise the speech within monologues and there are no interruptions, as opposed to interviews where there are frequent interruptions. The average time for transcribing an Arabic spoken file for a non-professional typist was calculated by adding the shortest time and the longest time taken in transcription and dividing it by 2. The result is 1:50:42. The transcribing speed was also calculated, i.e. the number of words per minute, and found the fastest to transcribe was the 'news reading'. Obviously, it was because words are pronounced clearly and slowly. The slowest is interviews which achieved only 4 words per minute.

As for processing the recordings, the files had to be edited to remove music. The files are annotated with the minimal demographic features for speakers such as sex, and occupation. Within the transcription itself places where there was an interruption or hesitation had to be marked using the standard elements that are used in English corpora such as <unclear>, <pause>, etc. It was problematic to mix Arabic and English in one text which is saved in Unicode. Once you switch into English script and type a letter the cursor jumps into another place in the document. Therefore, Arabic script was used to mark these elements.

It is worth mentioning here that obtaining spoken recording is not very easy as it requires personal contact to help with explanation about the purpose of the project and selecting the appropriate programmes. Some of the recordings obtained from Radio Qatar were full of music and lots of the programmes were not suitable.

We had contact with the director of the Arabic BBC in London in September 2003 by phone and had corresponded further by email and mail. It was promised that permission of copyright would be granted but never heard again. Later on there was a change of director and the process had to start again from the beginning. Knowing someone there was some help but it was too late to obtain the material and process the recordings. It was discovered later on that it would have been more practical if a meeting had been arranged with someone from the radio station in London and spend time selecting the appropriate programmes and more importantly speak to the authority about the purpose of the project. Even though some help was offered, the authority was still cautious about handing over enough recordings. Despite their agreement to provide us with recordings, they were not decisive about signing the letter of copyright.

The aim from the outset was to get recordings of everyday conversations and the best source was the Arab community here in the UK. This was a big project on its own and it requires a lot of time and contact with people. The original plan was to get speakers from different regions in the Arab countries but so far only small samples from male and female speakers of Qatar and Palestine were obtained.

Chapter 5

Result

After obtaining the results from the questionnaire and made decision on the form of Arabic and the varieties for the CCA, work on collecting the texts began only from the sites that gave permission. Section 5.1 reports on the result of compiling the CCA and the problems encountered. Section 5.2 lists the websites that granted permission to use, and reports on the difficulties found during the process of collecting the letters of permission.

5.1 The Corpus

In compiling the CCA, its internal structure was sought to match the needs of the users (teachers and language engineers) contacted by means of the questionnaire. The first step was searching for useful websites and obtaining permission of copyright. Fortunately, most of those contacted were pleased to use their material for teaching. Once copyright was grant text collection commenced⁴⁰. Every text is encoded with a header and the necessary information is added and saved as XML document. So far the CCA consists of 842,684 words in 416 files covering a number of categories. It must be pointed out that the list included in the questionnaire contained a mixture of text types and sources from which these text types are obtained. The sources are: newspapers, magazines, radio, TV and webpages. Table 5 shows the text categories which are derived from any of the sources, the number of texts in each category, and number of words.

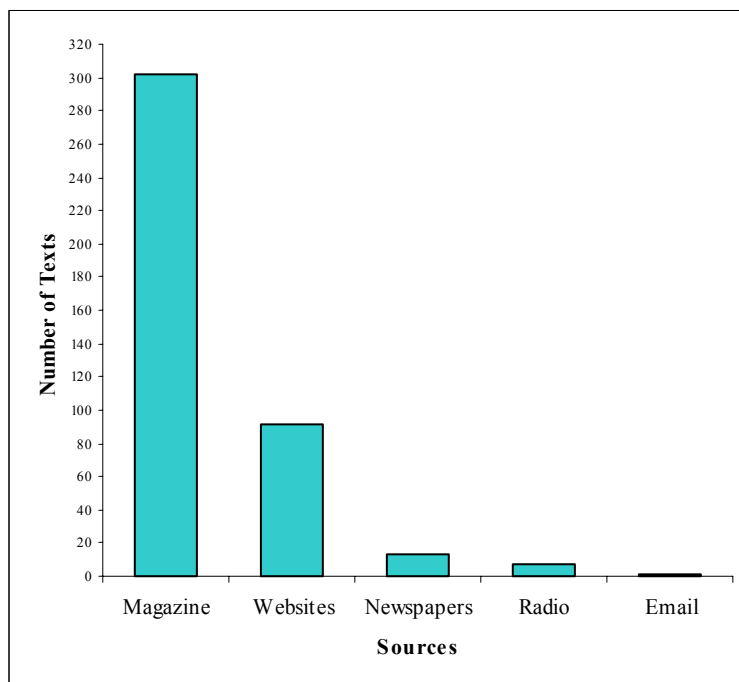
⁴⁰ There are some search engines that can be used for collecting a corpus such as WebCrawler, and Text Mining. In this instance, the texts were collected manually because only those sites who had given their consent were used.

Table 5: Number of texts and number of words in each category

	<i>Text Categories</i>	<i>No. of texts</i>	<i>No. of words</i>
Written			
1	Short stories	31	45,460
2	Television	Source	n/a
3	Education	10	25,574
4	Newspapers	Source	n/a
5	Radio	Source	n/a
6	Application forms		
7	Web pages	Source	n/a
8	Religion	19	111,199
9	Academic papers		
10	Business letters		
11	Advertisements		
12	Magazines	Source	n/a
13	Poetry	5	1,147
14	Formal letters		
15	Entertainments	2	4,014
16	Autobiography	73	153,459
17	Sociology	30	85,688
18	Conversation		
19	Tourist/travel	61	46,093
20	Instruction manuals		
21	Recipes	9	4,973
22	Geography		
23	Scientific documents	45	104,795
24	Emails	Source	n/a
25	Teen's stories		
26	Plays		
27	Restaurant menus		
28	Sports	3	8,290
30	Economics	29	67,478
31	Children's stories	27	21,958
32	Memos		
33	Fashion		
34	Health and medicine	32	40,480
35	Technical documents		
36	Financial documents		
37	User manuals		
38	Legal documents		
39	Internet computer documents	2	12,297
40	Calls for tender		
41	Patents		
42	Interviews	24	58,408
43	Politics	9	46,291
Spoken			
1	Education (MSA)	2	1,240
2	Sports (ESA)	3	1,736
3	Entertainment (colloquial)	1	1,377
4	Politics (MSA)	1	1,252

The original aim was to compile a million-word corpus. However, once in to the collection phase of the project, it became increasingly clear that this target was unrealistic considering the limited time and the difficulty of getting permission of copyright for the resources selected. Figure 11 is a graph showing the number of texts derived from the different sources. Most of the texts were obtained from magazines as they were the easiest to obtain copyright permission.

Figure 11: Number of texts used from the different sources



There are also some problems regarding text classifications, sample size, text grouping and representativeness. These will now be detailed.

Text classification

During encoding of a text type sometimes it is difficult to decide on which text category it belongs to and which domain. Sinclair (1996) examines in detail the problems of text classification and reports that corpus design makes use of some internal and external factors to decide on the text category. He points out that lots of text classification is based on topic as it is represented in newspapers and magazines. The Arabic corpora described at the beginning of the thesis seem to follow the topic criteria in their classification of texts. Although Sinclair believes that it is ‘a valuable feature of reflexivity of language.’ He states that ‘a typology based on such criteria will be untidy’. As a result he proposed 35 categories. However, Sharoff (2004) believes that such list is ‘too fine-grained’ and recommends another type of

classification which consists of only 8 main categories or general domains which include other types of texts. His proposed domains are (Sharoff 2004, p 1745):

- **NatSci** (math, biology, physics, chemistry, etc.)
- **ApplSci** (agriculture, medicine, ecology, engineering, computing, transport, etc.)
- **SocSci** (law, history, philosophy, psychology, language, education, etc.)
- **Politics** (inner, world)
- **Commerce** (finance, industry)
- **Life** (general domain eg. fiction, conversation, etc.)
- **Arts** (visual literature, architecture, performing)
- **Leisure** (sports, travel, entertainment, fashion, etc.)

The classification of the texts of the CCA will be based on Sharoff's as it seems to group a variety of text types under a general domain which is quite tidy. Table 6 shows a rough classification of the text types included in the CCA within Sharoff's general domains:

Table 6: CCA text types classified in Sharoff's domains

<i>Domains</i>	<i>Text types</i>
NatSci	Scientific doc.
ApplSci	Scientific doc., ecology, instruction manuals, geography, technical doc., user manuals, Internet comp.doc., health and medicine
SocSci	Education, academic papers, sociology, legal doc., religion,
Politics	Politics
Commerce	Business letters, financial doc., application forms, economics, call for tender, patents, memos
Life	Conversation, formal letters, interviews, advertisements, recipes, rest. menus
Arts	Poetry, short stories, children's stories, autobiography, plays
Leisure	Entertainments, tourist/travel, sports, fashion

Some of the text types in the above table can be classified under several domains depending on the topic that it handles. For example, 'interviews' can handle general topics but they can handle as well more specialised topics such as politics or medicine. The same applies to text types such as autobiography, memos, and patents.

Although it was intended to use contemporary texts, there are some books written by some well-known and prominent authors such as Taha Husain, Najeeb Mahfuuth, Tawfiq Al-Hakim, Jubran Khalil Jubran and others and which are not available on the Web. Despite that they are published in the 1960's or 1970's it was felt that works for such authors must be included for learners as they represent the best writing and thoughts of Arab scholars, especially as their books have been translated into many European languages. Foreign learners must be exposed to the Arabic versions.

Alongside collecting the texts we created a database file in Microsoft Excel which stores the ID number, title, source, number of words, year of publication, and author's name of each text. This database file is important to have for the organisation of the texts of the corpus and counting the words automatically.

Number of texts

As can be seen from table 5 above not all the text categories are obtained and this is due to the difficulty of finding resources on the Web for all the categories. It is difficult, for example, to get sources for business letters, menus, application forms, and plays. Some text categories are under-represented, i.e., they consist of a small number of texts. For example, the 'Children's stories' category has only a few samples of short stories from 'Al-Arabi Magazine'. As for spoken recordings from radio and TV, only a small sample was obtained.

Sample size

A specific size for each sample was not formally defined. However, the aim was to get the whole text of a document rather than excerpts. The majority of the collected texts were mainly short articles which consisted of a maximum 4000 words. Unfortunately, an electronic copy of a book could not be found. Texts that contained tables were avoided because when saving the file as a text the information in the table gets scattered. The texts collected were mostly published recently on the Web and written by a wide range of authors from all the different countries in the Arab world.

Text grouping

Some texts such as recipes and poems are very short reaching 100 words. Newspapers also contain short texts. It was not practical to have such short texts with their own header. Also this creates problem for concordancing. Therefore, it was decided to group several short texts into a single file with one header. This of course led to problems in encoding information in the header. For example, it was not possible to put all the names of the authors.

Representativeness

It was an important target to produce a well-balanced corpus in the sense of the selection of texts and number of words in each genre. However, problems of copyright permission or delay of responding of some sources such as science, Internet, computing, had prevented this goal from being achieved. In addition, it was extremely difficult to complete this project within the assigned one-year duration. Despite that a good number of authorisations were obtained for magazines, newspapers, and websites.

5.2 Copyright Issues

One of the important issues to consider when collecting texts for the CCA was to obtain copyright from the authors or owners of websites. The corpus would not be possible to use publicly unless permission was granted from owners of sites to use the material. Getting copyright is not an easy task. Several problems have been encountered: (a) finding the owners and contact address of websites, (b) Delay in getting replies, (c) online magazines display only the current issue, and there was no access for back issues.

However, the first stage in compiling the CCA was to identify some suitable sites and obtain email addresses as well as postal addresses, telephone numbers and fax numbers. Two letters had been prepared (see appendix V): one was to explain the purpose of the corpus and for the owners and authors to keep, and the other with a return slip for them to write their names and sign if they agree for their websites to be used. In case there was a delay in replying, a telephone call was made to confirm the arrival of the letter and give further explanation about the corpus and its purpose if needed. Most of the people who were contacted seemed to be cooperative. Some of them were very keen and they themselves called informing of the arrival of the letter and asking if any further help was needed (e.g. Al-Raijalaam newspaper, Kuwait). Others agreed but restricted the amount of the material to be used. For example, an owner of a site stipulated the use of only 10% of the material. Another requested some samples of how the texts are going to be represented their texts so that no one uses it for commercial purposes. Other owners requested taking material from all the sections but not a specific section, and this is related to matters of editing.

Table 7 is the list of addresses of resources for which permission of copyright had been received.

Table 7: Sources with copyright permission

<i>Address of sites</i>	<i>Source type</i>	<i>Method of contact</i>	<i>Method of replies</i>
http://www.un.org/Arabic/ http://www.un.org/English/ The secretary of the Publications Board, United Nations New York, NY, 10017, USA	Website	Email and mail	Email
http://www.alarabimag.com Ministry of Information PO Box 748 Safat-Kuwait Editor: Dr Suleeman Ibrahim Al-Askari Tel: 00965-2437875 Fax: 00965-2437877	Magazine	Mail, fax and telephone	Express mail
http://www.ofouq.com Editor: Mr Mohammed Al-Nabhan 410-1129 Meadowlands Dr. Ottawa, ONT. K2E 6J6 Canada	Magazine	Email and mail	Email and mail
http://www.alhourriah.org Editor: mu'tasim Hamada PO Box 11488 Damascus-Syria	Magazine	Mail	Mail
http://www.arabicstory.net Person in charge: Mr Jubair Al-muleehan Email: Jubair22@hotmail.com	Website	Email	Email only
http://www.akhbarelyom.org/akhsaa/ Editor in chief: Mr Mohammad Barakat Address: Dar Akhbar ilyom 6 As-sahafa st. Cairo-Egypt Tel: 5782600-5782500	Magazine	Mail	Email only
http://www.lahaonline.com PO Box 286083 Riyadh 11323 Saudi Arabia Tel: 0096612791300	Magazine	Mail	Mail
http://www.sayidaty.net Editor in chief: Hani Naqshabandi Arab Press House 184 High Holborn, London WC1V 7AP Tel: 02078318181 Fax: 02074301280	Website of 'Sayyidaty' magazine	Mail	Mail

Email: sayidaty@hhsaudi.com			
http://www.ecoworld-mag.com Editor in chief: Dr AbdelAziz Ismail Daghistani Tel: 055-479-479 Email: dradaghistani@yahoo.com PO Box 1661 Riad 11441 Saudi Arabia	Magazine	Mail	Mail
http://www.almarefah.com/ Editor in chief:- P O Box 7 Riyadh 11321 Tel:419-4040 Fax:419-4747 Free Fax: 800-124-2277	Magazine	Mail	Mail
http://www.arabcomputing.com/ Editor-in-chief: Khalid Hamadeh Email:info@arabcomputing.com 297 Preston New Road Blackburn, BB2 6PL UK	Magazine	Mail	Mail
http://arabmedmag.com Editor in chief: Dr Mazin Al-loujami Email:loujami@net.sy PO BOX 36164 Damascus, Syria	Magazine	Mail and email	Mail and email
http://aklaat.com/ Abu Dhabi P.O.Box 51526 UAE 4430496 Mr. Abdelraheem Al-Shaikh Email:chief@aklaat.com	Website	Mail	Mail
http://www.islamonline.net PO Box 22212 Doha-Qatar Tel 00974-4457744 Fax: 00974-4358844 Or public relations editor Ghada Tel:(mobile) 0020106133103/(office in Egypt) 2023380337	Website	Mail	Fax
http://www.alraialaam.com Editor: Jassim Marzuuq buudi Email:editor@boodai.com Address: PO Box 761 Safat 13008 Kuwait Fax:00965-4838352	Newspaper	Email and mail	Email and waiting for the mail
http://www.kisr.edu.kw/science/	Magazine	Mail	Mail

PO Box 24885 Safat-Kuwait 13109 Tel 4818630			
http://www.nizwa.com/ Editor in chief: Saif Al-Rahbi Tel: (00968) 601608 Fax: (00968) 694254 Email: saif@alrahbi.com PO Box 855 117 Al-Wadi Al-Kabir Saltanat Oman	Magazine	Mail	Mail
http://www.raya.com Al-Rayah Newspaper Editor: Email: edit@raya.com Fax: +4371353	Newspaper	Email and mail	Email

Table 8 is a list of addresses who were contacted but still awaiting permission.

Table 8: Sources awaiting copyright permission

<i>Address of sites</i>	<i>Source type</i>	<i>Method of contact</i>
http://radioqatar.com/ Head: Mr Abel-Rahman Nassir Al-Obaidan	Radio	Mail
http://www.alamalcomputer.com 21 A Emirat Elebour, Salah Salim, Heliopolice, Cairo, Egypt Tel: 202-4022816 Fax: 202-4022816 Mobile: 20105430656 Email: hisham@alamelcomputer.com or info@alamalcomputer.com	Magazine	Mail
http://www.anfy.com Fabio Ciucci Via P. Paolini, 247 55100 Lucca Italy	Website	Mail
http://www.bhaintv.com Executive chief: Khalil Ibrahim Al-Thawwadi Tel: 973-686000 Fax: 973-681544		
http://www.alwasatnews.com/ Dr Mansur Al-Jamri PO Box 31110 Kingdom of Bahrain Tel: 973-17596999	Newspaper	Mail

Fax: +973-17596900 Email: letters@alwasatnews.com		
http://www.akhbar-alkhaleej.com/ Editor-in-chief: Anwar Mohammed Abdelrahman P O Box 5300 Manama-Bahrain Tel: 620111 Fax: 621566 Email: info@akhbar-alkhaleej.com	Newspaper	Mail
http://www.BBCArabic.com Editor of online BBC: Husam Al-Sukkari Tel: 02075572525 Mobile: 07808723893 Email: Hosam.sokkari@bbc.co.uk Arabic BBC Room 418 CB Bush House, Strand, London WC2B 4PH Head of the Arabic section:Mr Mustafa Anwar: Tel:00442072403456	Website	Mail
http://www.acookweb.com PO Box 105816 Riyadh 11656 Saudi Arabia Fax 00966-1234-6169	Website	Mail
http://www.bahraintv.com Executive chief: Khalil Ibrahim Al-Thawwadi Tel: 973-686000 Fax: 973-681544	TV	Mail
http://www.alayam.com Editor in chief: Mr Isaa al-Shaiji Email: al-shaiji@alayam.com Al-ayam newspaper PO Box 3232 Manama, Bahrain	Newspaper	Mail
http://www.al-jazeera.net Manager: Mohammed Dawood Tel: 00974-4382-803 (direct), 00974-4382-777 Email: m.dawood@aljazeera.net Manager: Ibrahim Hilal (TV) Tel:009744-890827	Website	Telephone and mail

Despite the difficulty in getting copyright permission, nineteen owners of websites agreed. This was a good number of sources considering the time constraints. Writing letters and contacting people was time consuming and sometimes it was difficult to focus on the essential work.

5.3 Summary

Since compiling the CCA was limited by time and obtaining permission of resources, it was not possible to collect a balanced number of texts in each category. One of the significant issues was including more spoken and colloquial material, but this turned out to be very difficult in terms of obtaining it and inputting it. However, given more time the results would have been better. Thus, for this corpus to be more balanced and to fit in well with the requirements of the language teachers, resources for all the other categories that have not been collected will need to be found, and to add more spoken data from radio and TV of different regions in the Arab countries. The possible solution would be to use a speech-to-text program to handle the spoken part of the corpus. If a program that converts speech into text fairly accurately can be found, it would be a great advantage.

Chapter 6

Uses of the Corpus

The corpus is a resource to use for studying linguistic phenomena as well as for teaching languages. In order to do that, some special tools called concordancers which do the task of searching, sorting, and classifying have been designed to help us manipulate the data. Nowadays several programs are available. Some are commercial such as WordSmith, MonoConc and ParaConc. Others are free such as ConcApp⁴¹ and Wconcord⁴². These tools work perfectly well on English and other languages with Roman script but until now there is no efficient tool available for processing Arabic. For this reason, corpora with Roman script have already been used in the field of language teaching. Section 6.1 illustrates the application of corpora and concordancers for teaching languages other than Arabic. Section 6.2 shows what the present programs are capable of doing for Arabic texts hoping that this will encourage some experts to develop a program that works better.

6.1 Using Corpora for Language Teaching

With the recent growth in the number of corpora, especially that many of them are freely accessible, many teachers came to recognise the importance of using authentic data for teaching. One of the common activities used is concordancing. It is a technique through which learners can search and sort the data to elicit certain types of information. It has been pointed out (Johns 1988) that the use of concordancing is a valuable activity for a number of factors. Learners have control over their learning process. They became active and responsible and the teacher became a guide rather than the source of all knowledge. They have access to large natural texts instead of made up sentences, and learners arrive at the rules of the language by searching and discovering for themselves instead of being told the rules ahead. This deductive method is important in learning a new language. When the students arrive at the rule by discovery procedure it is easy for them to remember it. This approach is referred to in the literature as Data Driven Learning (DDL). A concordance program became an essential tool for searching as it saves time and presents the data very neatly to the learners. Several studies argue for the value of corpora and concordancers in the teaching of languages.

⁴¹ Available at <http://vlc.polyu.edu.hk/pub/concapp/concapp.htm>.

⁴² Available at <http://www.pef.zcu.cz/andy/martinek/wconcrd>.

Minugh (1997) points out the difficulty of teaching English in a country that uses another language (Sweden, in his case). It is hard sometimes to give judgements on the acceptability of an expression if you do not use the language often. In such circumstances he recommends that corpora provide an immense help to teachers. He specifically refers to newspaper CD-ROMs and suggests a number of ideas for teaching language. They can be used for searching the meaning of new words or phrases. The word *dweeb* for example has no mention in several modern dictionaries but when checking the word in the New York Times corpus he found eight occurrences of it. They can be also used to check the current usage of a word. He searched for the current meaning of the word *fondle* in four dictionaries and found that it means 'to touch gently and lovingly'. However, when checking the *Independent* CD-ROM for 1992 he found 33 uses for this verb. Of the 33 instances only 3 gave the dictionary meaning. The rest gave other meanings. Corpora can also be used for investigating grammatical constructions. One example he mentioned is the type of constructions that follow the expression '*it is high time...*'

In Swedish grammars two forms are given: *for*+infinitive, and *that*+subjunctive. But on searching the New York Times CD-ROM more complicated structures were found. In addition, corpora can be a source for learning new vocabulary and expressions of certain registers and styles. Minugh concludes that 'of particular interest is their enormous potential for advanced learners' (p78).

Dodd (1997) highlights the uses of corpora for advanced learners in three areas: as a general resource for students to browse through, as a resource for discovering grammatical rules of the language and as a resource for students' projects. Simple concordance files offer an opportunity for students to build their vocabulary and increase their knowledge of word formation. Information on frequency provides clues for translation. For example, to find the correct translation in German for the word 'completely', the frequency file based on BZK (Bonn Newspaper Corpus) offers three different frequencies: 24 of *total*, 70 of *absolut* and 409 of *völlig*. This gives students clues on the use of the correct form based on frequency of usage and context rather than guessing the word. Corpora also have applications in advanced grammar courses. Students can gather information about a particular grammatical rule then they search the corpus for examples to match the rule. This is an interesting exercise as they might discover examples that are not covered by this rule. This gives them a chance for forming new insights into the structure of the language. Students in MA level can use the corpus to work on their projects and this proves to be interesting and challenging to existing theories about the language.

Cobb, Greeves, and Horst (2001) detailed a case study to show the effect of using authentic material and on-line resources (referred to as R-READ) on the acquisition of vocabulary of second language learners. The program they used is located on the Web⁴³. Learners can read and/or listen to a novel and they can click on any difficult word and get a concordance. They also can click on on-line dictionary and record meaning of words in a personal database. A French parallel site is also available. Two second language learners were compared: R, an intermediate German learner who did not use online resources and the other, J, an intermediate French learner who used online resources. The comparison between their rates of vocabulary acquisition confirms the effectiveness of the R-READ approach. Both readers scored 45 per cent of their target words unknown in texts. At the end R decreased his unknown words by only 7 per cent but J decreased it by 38 per cent. In the known category R did not increase his list while J increased it by 250 percent. Moreover, on the translation post-test J produced correct answers after three readings, while R after 10 readings. This confirms the value of authentic material and online resources in building the second learners vocabulary.

The general view is that the activity of concordancing and analysing the result is more suitable for upper intermediate and advanced learners but not for beginners. A study by Hadley (2001) proves that corpora can benefit beginners and low intermediate learners of English. In the Nagaoka National College of Technology in Japan the technique of concordancing (based on Nagaoka Kosen Corpus) has been used along with the textbook. The result was that there was an improvement in the writing skills and test scores of most of the learners. However, some problems were encountered. These are related to the amount of data and the complexity of the structure and vocabulary. Some students found it difficult while others found it challenging. Despite these problems, on the whole it was more interesting for them to discover the rules of grammar than be told.

Although it has been widely reported that corpora and concordancers are useful in computer-assisted language learning, there are not many studies that test the validity of this claim in comparison to traditional methods. However, two studies, one by Stevens (1991) and one by Cobb (1997) did seek to compare the approaches. The first is a controlled experiment conducted at Sultan Qaboos University in the Sultanate of Oman. Stevens' experimental task was to have students recall a known word to fill a gap in a text, either a gapped sentence or a set of gapped concordance lines for a single word. The subjects were male and female first year students studying science. The experiment was conducted in two sessions using a concordance-based exercise and the other the gap-filler. Some practise has been given to

⁴³ See <http://132.208.224.131/callwild/>.

students prior to the experiment so that they become familiar with the concordance format. The overall result shows that students did better in the concordanced based exercise than the other type. The learners would retrieve a word from memory more successfully when cued by the concordance lines, in spite of their fragmentary nature.

The study by Cobb is also conducted in the same institution and tested the hypothesis that ‘a computer concordance might stimulate potentially rationalize off-line vocabulary acquisition by presenting new words in several contexts.’ To test this hypothesis he developed an experimental lexical tutor to introduce new words (20 words per week) to subjects either by means of concordances or other sources. The subjects were more than 100 students learning English in the first year and the experiment lasted for 12 weeks. Five types of activities were used: choosing a definition for the concordanced word, word recognition by identifying the correct word that fills that concordanced lines, spelling words, choosing words for new texts, and writing words for new texts. Students were tested before and after the experiment to measure their word knowledge. It was found that they achieved a 21.5% higher score when using concordancing over not using concordancing. As for the weekly quizzes it was found that the mean score without concordancing was 63.9% and with concordancing was 75.9%. This supports the idea that use of corpora and concordancing has a positive effect on the students’ knowledge of vocabulary.

6.2 Using Corpora for Teaching Arabic

In the previous section the use of corpora and concordancers in the teaching of English and other languages was demonstrated. It is believed that the use of such resources and tools are very important to use for teaching Arabic for foreign learners. With the use of a concordance the corpus would be a source of knowing meaning of neologisms. Lots of words enter the language and they are too recent to be found in the existing dictionaries. This creates a problem especially for students learning translation (from English into Arabic) at advanced levels. A good way to solve this problem is by using a corpus and analysing the context of the word. Some terms go out of fashion and get replaced by new words.

Unfortunately, the use of such resources in teaching Arabic is very limited. The only case known is that of the English Department in Kuwait University where a parallel corpus (Arabic and English) is used (Al-Ajmi 2003). It is used for teaching lexicography and translation courses using Al-Idrisi search program which has been developed by Sakhr Software. The corpus developer provided special, albeit limited, access to this resource. Figure 12 is an example of searching for the word (short). It shows the occurrence of the word in both

the English and the Arabic texts, and you can click on 'More' to read more of the context in which the word occurs. You can get the same meaning for the word, the antonym, the word with affixes and without. It produces good results and both the English and Arabic texts are correctly matching.

Figure 12: The concordance of قصير (short) in a parallel corpus using Al-Idrisi search tool

نظوي نتائج البحث على 4			
الموضوع :	كل الموضوعات	الكتاب :	الذكاء العاطفي
النص الحر :	قصير		
النص العربي	English Text		
نضرب مثلا بالتجربة التي شاهد فيها المتطوعون لقطات من فيلم قصير المزيد ...	For example in one experiment volunteers watched short film clips More...		
والنفس القصير المزيد ...	shortness of breath More...		
على الأكل على المدى القصير المزيد ...	at least in the short term More...		
إذ إن الشعور الجميل على المدى القصير يقابله انهيار حياتهم المزيد ...	a short-term good feeling in exchange for the steady melt-down of one's life More...		

In this section concordancing an Arabic text using Monoconc program will be demonstrated. The data used is a 'raw' corpus from the CCA, and the file is created in Windows 2000 English version, with the text encoded in Windows-1256 (Arabic), rather than the usual UTF-8 which is not supported by MonoConc. In this program there is no option for the Arabic language among the languages listed, but the Arabic text can be displayed by choosing any Arabic font available. The result is shown in figure 13:

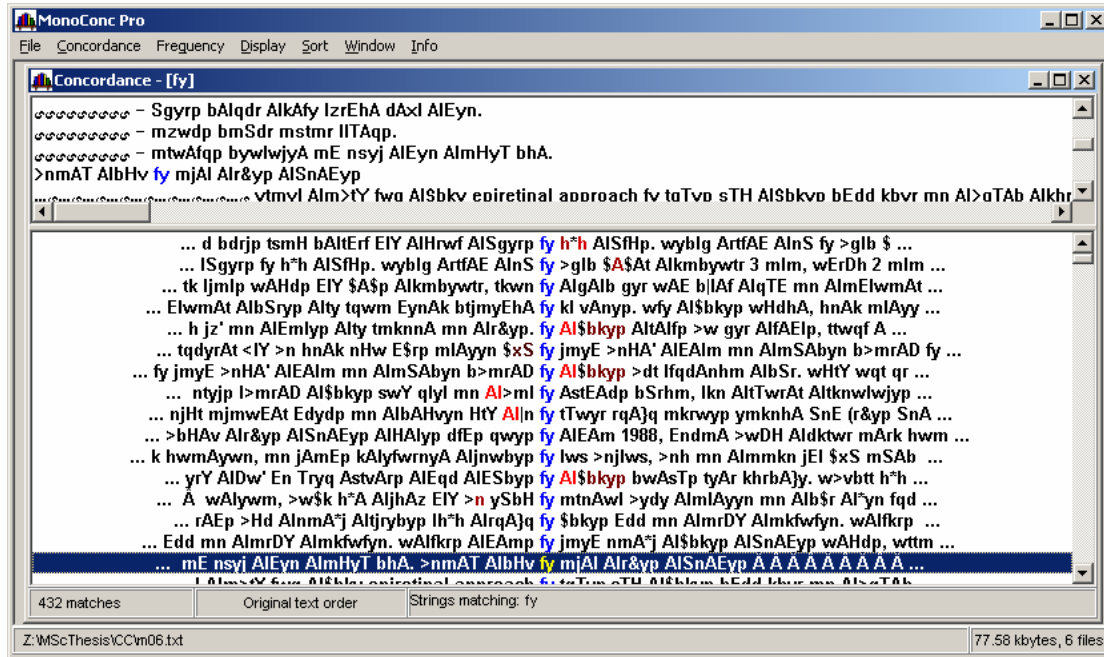
Figure 13: Concordance of البلد (the country) using Monoconc in English Window.



The above screenshot shows that although the program displays the target word properly aligned but the order of the words in the sentence is not correct, and when clicking on the target word to view its full context in the window above, the text can not be displayed in Arabic.

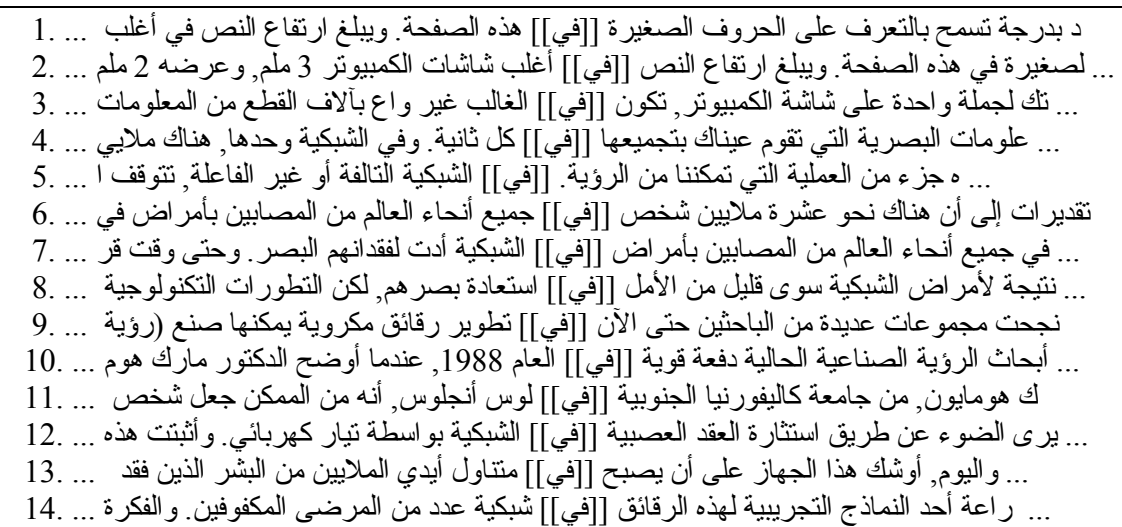
Since this program works better on languages that use Roman characters, transferring the Arabic text into Roman characters might give a better result. In order to do that, a certain procedure had to be followed to create a concordance sample. A transliterating program, which was developed by a colleague here at the University of Leeds, was used. This program transfers the Arabic text into Roman characters using Xerox–Buckwalter’s transliterating system (Beesley 1997, 1998) (see appendix II). This system allows for a one-to-one correspondence between the Arabic character set and the Roman character set. After transferring the text into Roman characters Monoconc was then used to conduct any type of search. Figure 14 shows the concordance of the preposition fy /fi:/ (in).

Figure 14: Concordance for the preposition 'in' in the transliterated text



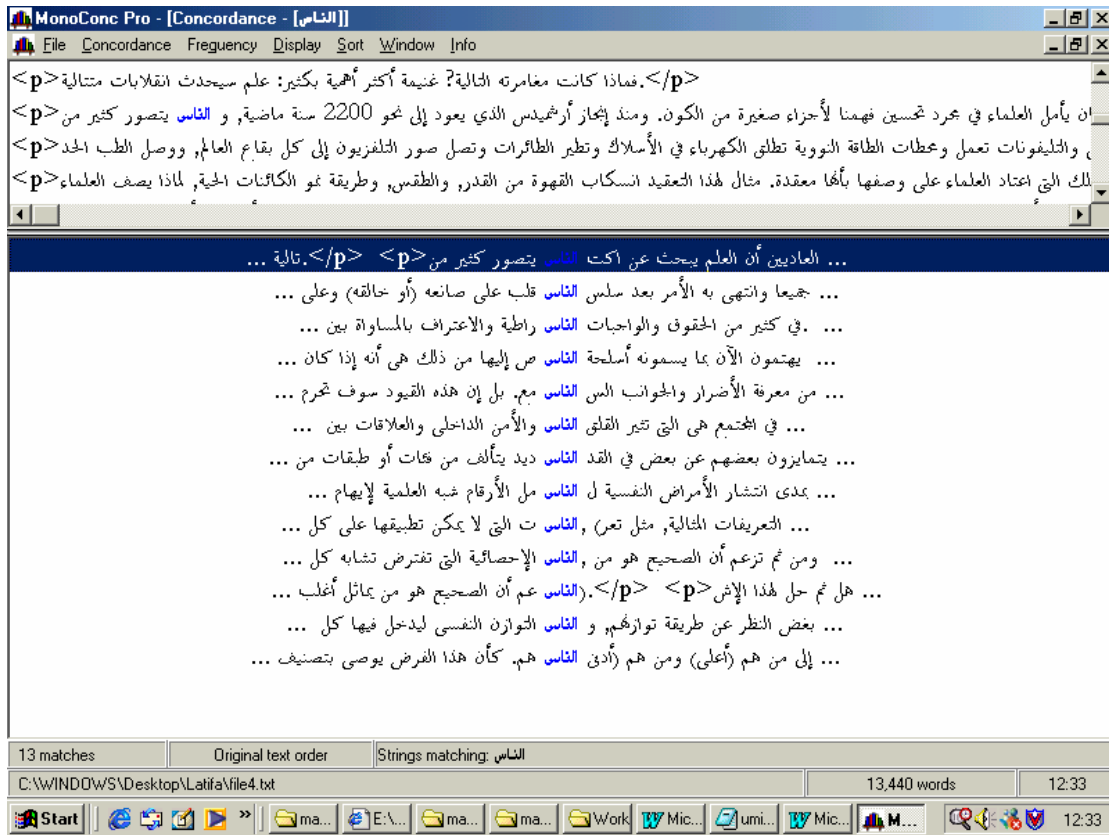
The next stage is to save the results as a text file and then transliterate back into Arabic. The result can be then viewed, as shown in figure 15.

Figure 15: Concordance for the preposition 'in' after transliterating it into Arabic



The same process of concordancing is followed using WordSmith. However, Monoconc and Paraconc work better on Arabic when used in Arabic Windows. There is no need for transliterating as it works directly on Arabic texts, as illustrated in figure 16.

Figure 16: Screenshot of the concordance of الناس (people) using Monoconc in Arabic Windows



The problem with this concordance is that the order of the words is not correct. Reading Arabic should start from the first word on the right. But in this concordance you have to read the string of words that occur after the target word then back to the target word then the first word on the right. The example below shows the order of reading.

في كثير من الحقوق والواجبات	الناس	والإعتراف بالمساواة بين
3	2	1

Compare this with the English example below if you have to read it in the same order.

licked the milk	cat	the starving
3	2	1

In order to solve the problem of word order, we have to save the concordance file as text and then open it again.

Figure 17: Concordance of الناس (people) saved as text

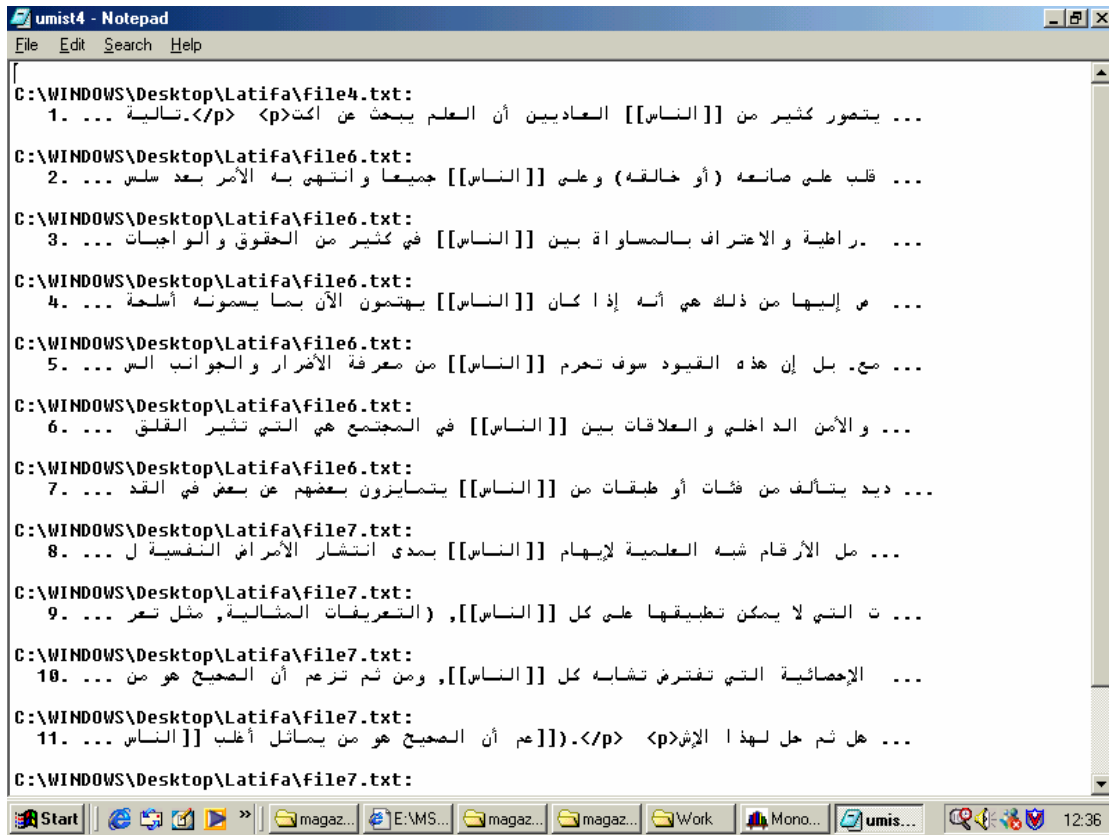
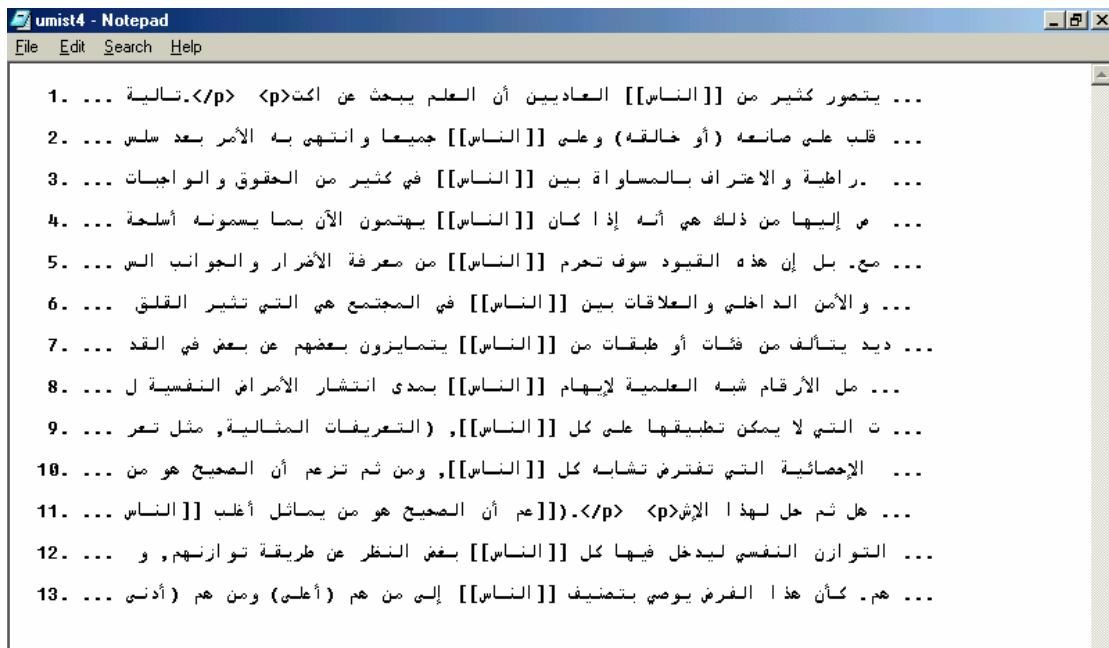


Figure 17 shows that the order of words is correct, although extra lines were automatically added which can be deleted manually or by using Find and Replace command. Figure 18 is the final stage of concordancing an Arabic file:

Figure 18: Final stage of concordancing of الناس (people)



Performing concordancing on the CCA leads to problems. The files are saved in XML format which Monoconc does not handle. Also, the files appear with strange symbols. This is probably because of the difference of fonts. However, a workaround was achieved by copying and pasting the content from the CCA files and saving them as text files in Arabic Windows. It is not fully understood why this problem occurred (nor why the solution worked!)

As you may have gathered from the prior discussion and the illustrations presented, concordancing of Arabic is not efficient at the present time. Those who are using concordancing for Arabic texts are basically using Arabic Windows and by saving the file as text they can get a list of key word in context (KWIC) but with some problems which require them to do some editing. For example, the lexicographer, Hoogland (2003), reports several problems of the use of Monoconc. One of the serious issues relates to the order of reading the displayed concordance as already illustrated. Although for him it became normal after several uses but it is still a problem to use for learners whose knowledge of Arabic is not of high level and this could create misreading of sentences. Other problems reported are related to the fact that Arabic texts are not vowelised and add to it the use of affixes and infixes which produce ambiguity in the concordancing process. The option of wild cards can be used but, this results in getting unwanted occurrences especially the occurrence of weak roots where the second or third radical is hamza, waw or ya and also those that have prepositions and conjunctions connected to them. So in the search of a certain root in the corpus it is possible to get all the occurrences of three strong radicals but sometimes when the first letter is a conjunction like *wa* which is normally connected to the verb it creates a problem. Only 37% of all roots can be correctly processed.

Despite the complication of the use of Monoconc with Arabic text it was a useful tool for producing frequency lists, identifying frequent spellings and finding collocations which are important for developing a dictionary. Handling large data manually is an impossible task so any analysis tool could prove to be practical to use than nothing. What is needed is a concordancer that can support Unicode encoded files (so that it can read Arabic texts), and to work properly with Arabic, i.e., correct order of words and no need for transliteration or post-editing.

6.3 Development of a New Concordancer for Arabic

At the end of this project, a presentation about the state of the art of Arabic corpora and the processing tools available for use was given in the School of Computing at the University of Leeds. The presentation not only created an interesting discussion and exchange of useful

ideas among the researchers, but also prompted one of the young researchers, Andrew Roberts, who is non-Arab and does not speak Arabic, to develop a basic concordancer for Arabic in a very short time. The concordancer is referred to as aConCorde. It was developed using the Java programming language and was designed with internationalisation features so that it can be easily extended in the future to suit other languages. One of the advantages of using Java is that internally, programs represent text in Unicode (UTF-16) rather than ASCII. In addition, it gives support for other encodings such as UTF-8 and ASCII. Its graphical user interface components were also designed to be able to display right-to-left languages. Since Java provided many facilities to process and display Arabic, it was easy to use it to create a concordancer. Not only the underlying algorithms work for Arabic but would work for other languages such as Hebrew, Japanese and many other languages. The other advantage of Java programs is that it is multi-platform. As long as the Java Runtime Environment (JRE) is installed on the system, it can run a Java application. aConCorde was written on a Linux environment, but it runs exactly the same on Windows and Apple Macs, Solaris, and other Unix-based operating systems.

Advantages

- Displays concordance correctly and target word is well aligned (See the screen shot below in figure 19).
- English or Arabic user interface.
- Currently can cope with UTF-16, UTF8 and ASCII encoded files, as well as other dedicated Arabic character sets.
- Can run correctly on many operating systems (including non-Arabic Windows).
- Can be easily extended to work for other languages.
- Open source - program is freely available to anyone (as will the source code).

Disadvantages

Since the program is developed very recently, it has some disadvantages:

- Contains only basic functions in comparison to the well-developed programs such as WordSmith and MonoConc.
- Does not recognise XML markup or any form of markup language.

Figure 19: Concordancing of the word **غرناطة** (Granada) with word frequency

The screenshot shows the aConCorde software interface. The main window displays concordances for the word 'غرناطة'. The concordances are listed in a table with columns for 'تكرار' (Frequency) and 'كلمة' (Word). The frequency table shows the following data:

تكرار	كلمة
1	14
1	1492
1	18
1	1802
1	1829
1	1898
1	1933
1	1936
1	1975
1	1998
1	20
1	2003
2	500
1	600
1	700
1	800
1	أثار
10	أخر
1	الأف
2	أه
1	أبحث
2	أبراج
1	أبيض
2	أبو

The concordances list shows the following text snippets:

- الذي حاول ان يحيط بروح **غرناطة** بما فيها من ناس وبيوت
- إلى بحر التاريخ وكوابيسه, كانت **غرناطة** تأتي إلا أن تسقيننا من
- عنها الشاعر الإسباني لوركا: (إن **غرناطة** تقف على جبلها وحيدة منعزلة.
- على إقامة هذه المؤسسة في **غرناطة** حتى تكون قلب الإسلام في
- حفايق عظمي, وربما تعطينا حالة **غرناطة** علامة علي ما تعنيه السيدة
- كان علينا أن نعود إلى **غرناطة** في اليوم المحدد لافتتاح المسجد.
- روبو: أمير الجماعة الإسلامية في **غرناطة** قال لي: (من وجهة نظر
- توقع معاهدة استسلامها, تجرّع أهل **غرناطة** كأس السم كاملا بعد سقوط
- من فوق تلالها. رحلتنا إلى **غرناطة** كانت مختلفة, في البداية صنعنا
- طارت به الريح في مياه **غرناطة** لا مجاديف سوى الرفرات أه
- أنفسهم, بعد أن طرد أهل **غرناطة** لجا الجزء الأكبر منهم إلى
- إلا أنهم عند انسحابهم من **غرناطة** لم ينسوا أن ينسفوا برجين
- أبو عبد الله الصغير آخر ملوك **غرناطة** من أحد الأبواب, حاملا معه
- كان السلطان الذي كون مملكة **غرناطة** من فتات الممالك الهلوية قد
- أقدم الكليات التي انشئت بجامعة **غرناطة** منذ القرن الثامن عشر, وتهتم
- مشاعره: اسم الله يتردد في **غرناطة** من كان يصدق أن شمس

The status bar at the bottom indicates: 5 = حجم سياتي النص | (ACorpus.txt) UTF8 | ملف المدونة: 5151 = مجموع الكلمات 2625 = الكلمات المجردة

The work on this concordancer is still going on but since the developer is occupied with his own research, he would hope to have a team of developers working with him to speed up the progress on this program. He also aims to reduce the disadvantages of the program such as adding the facility of filtering out the XML markup and ensuring a robust user interface and adding major new features for language analysis. This concordancer is freely accessible on the Web⁴⁴.

Only recently during the Teaching and Language Corpora conference (Talc06) that the general purpose search tool Xara (or Xaira) has been tested with the CCA (Bernard and Dodd 2003, Bernard 2004). It seemed to work well and perform basic queries. However, there was no chance to test it with more complex queries.

⁴⁴ It is available at <http://www.comp.leeds.ac.uk/andyr/software/aconcorde>.

Chapter 7

Discussion and Conclusion

This chapter discusses the results of the research: designing and developing a corpus of contemporary Arabic, states the contributions and future development of the corpus. Section 7.1 discusses the results, and sections 7.2 and 7.3 present the contributions and future development of the corpus.

7.1 Discussion

The main purpose of this research was to design and develop a free corpus for teaching Arabic as a foreign language. Handling this research at this time is much easier than in the 1980's or early 1990's because Arabic texts are widely available on the Web. Also, what made this project easy to pursue was the support received from some owners of Web sources who were enthusiastic about the idea of spreading Arabic for assisting foreigner learners.

Although the ambitious aim was to fully develop the CCA, it was unrealistic to achieve that considering the limited time. A number of texts for several categories had been collected, but it was hard to have access to the other categories. This is due to several factors:

- Lack of enough sources in the field.
- Absence of resources (business letters, advertisements, plays).
- Difficulty in finding the owner and sometimes the address of the website.
- Not having enough material on the site, especially that some magazines do not make back issues available.
- Having several owners of copyright for the same website where one agrees and the other does not.
- Having a site where its owners are multiple and so it is difficult to contact each one such as the case of Arabic BBC.

However, despite these difficulties, 842,684 words in 416 files had been collected. At the initial stage, a large number of texts from some reputable general magazines such as 'Al-Arabi' had been collected. This is because the texts are of reasonable length and the articles

cover interesting topics which attract a wide range of audience. This speeds up the growth of the corpus. Texts from other sources had been collected as well.

Figure 20 summarises the result of the research and shows a comparison between the different categories in the number of texts.

Figure 20: Number of texts in each category

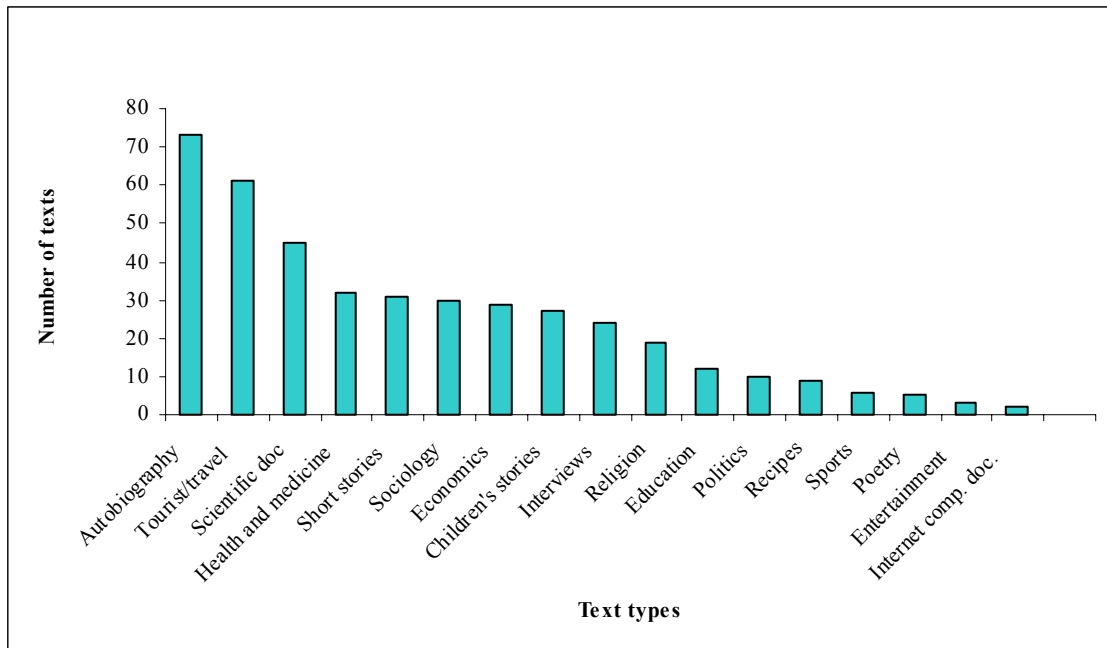
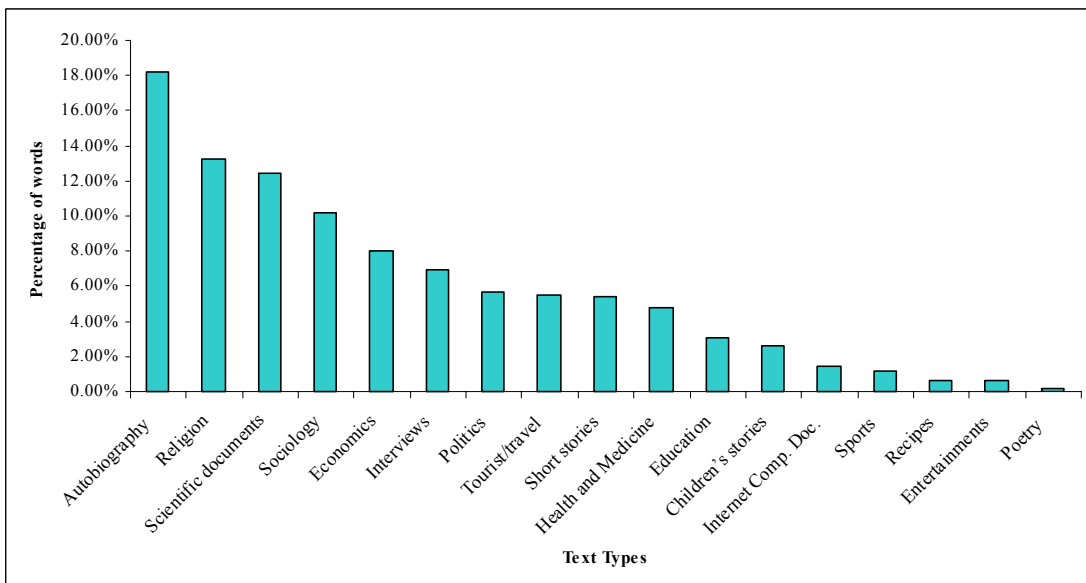


Figure 21 shows that number of words in each category does not depend on the number of texts.

Figure 21: Percentage of words in each category



The general aim was to develop a representative corpus of Arabic and to match the results of the survey obtained from teachers and language engineers. The results shown in the above figures show that this aim has not yet achieved. This is mainly due to time constraint and problems encountered in text encoding in the early stage of the project and delay in getting copyright permission.

The genres that are identified by language teachers to be most important are:

Category 1: short stories, TV, education, newspapers, radio, application forms, religion and web pages.

Category 2: academic papers, business letters, advertisements, magazines, poetry, formal letters, entertainments, autobiography, and sociology.

And genres that are identified by language engineers to be most important are:

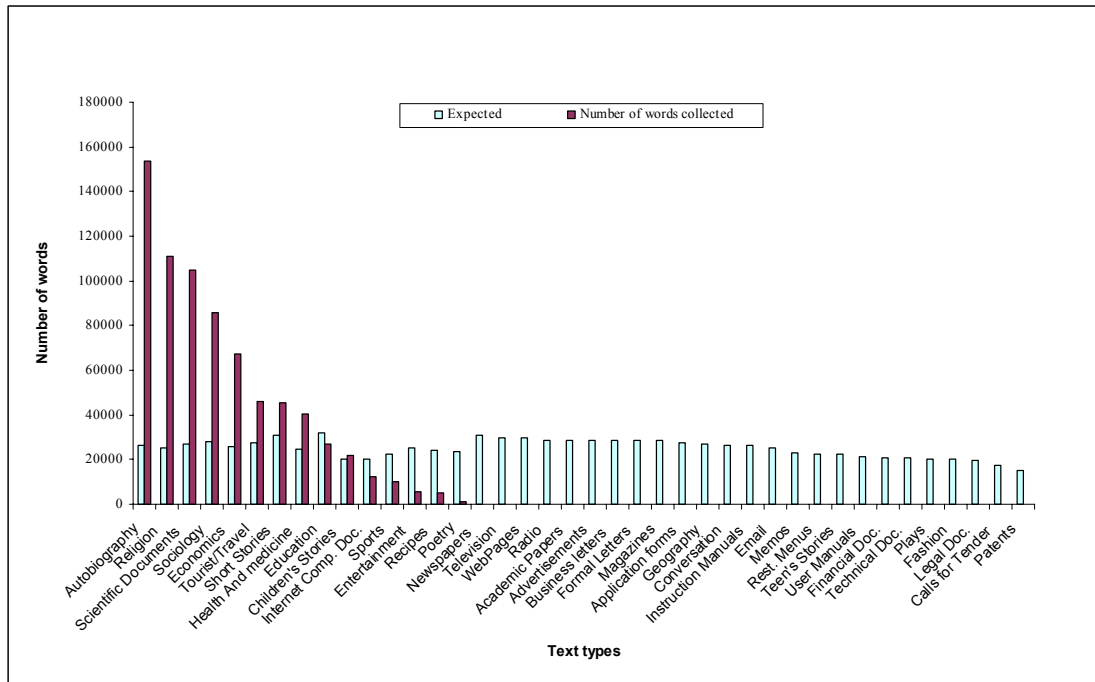
Category 1: newspapers, education, sociology, geography, scientific documents, economics, health and medicine.

Category 2: web pages, formal letters, academic papers, business letters, advertisements, magazines, tourist/travel, user manuals, Internet computer documents, short stories.

Based on this result it was expected to have collected different number of words in each category. In chapter 3, figure 6 shows the number of words for each category that should have been collected. Figure 22 shows how many words had been collected in comparison to the expected size for each category. Table 9 summarises the result in terms of the relative difference between actual and expected word acquisition.

Table 9: Types of the categories that match or not match the target size of the CCA

On target ($\pm 5\%$)		-
Over	5-15%	Children's Stories
	>15%	Short stories, tourist/travel, health and medicine, economics, sociology, scientific documents., religion, autobiography
Under	5-15%	-
	>15%	Education, Internet comp. documents, sports, entertainment, recipes, poetry

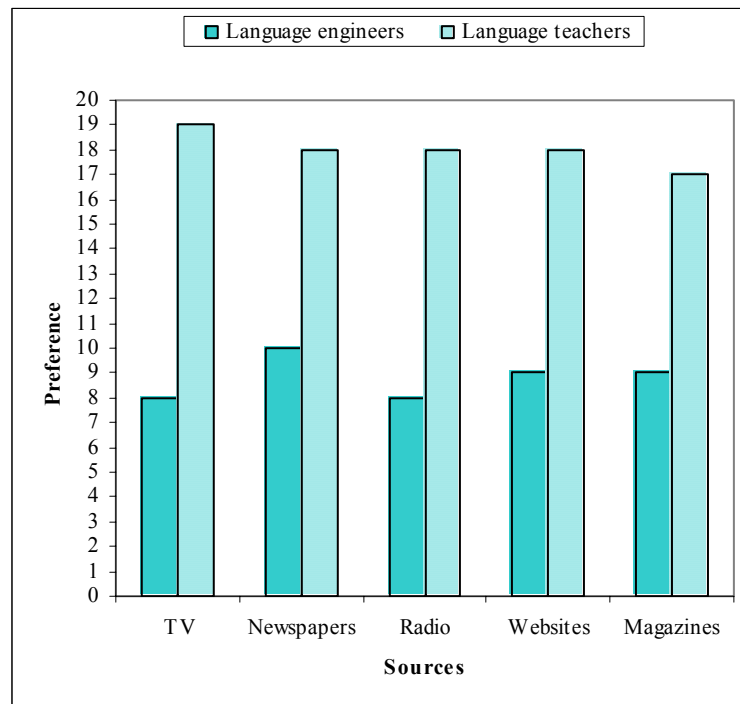
Figure 22: Comparison between the number of words collected and expected to collect

Some of the other categories such as business letters and formal letters are not easy to get from the Internet. One way of obtaining these kinds of genres is by contacting companies and getting hard copies which then need to be scanned. As for other texts such as advertisements, transmissions from radio stations contain spoken adverts which would be useful for learners, however, permission would have to be granted from the advertiser, rather than the radio station. An inquiry had been made about using such material but not yet known if it is possible to get copyright. Specialised resources for other categories, such as geography, were not found. It was also difficult to find websites for obtaining application forms. The application forms are either stored as PDF files or online application forms. Arabic websites about the Internet and computing are very limited. Those found were not easy to get copyright permission from, e.g. the Arabic PC magazine. For academic papers, no attempt was made to find out sources for them but it might be easy to obtain them from Arabic journals. Newspapers are widely available and so far two sources had been obtained to use from Kuwait and Qatar. Categories such as poetry are available, however, encoding them takes longer than normal texts, yet they contain fewer of words which does not contribute much to the size of the corpus. As for radio, an attempt had been made to get some recordings from outside the UK but was not successful in getting enough and in the range wanted. It is important to have a large number of spoken texts in the corpus especially that it is aimed for learners of Arabic whose priority is to use it for communicating.

Out of the six resources from which the texts were obtained (magazines, websites, radio, TV, newspapers and email), all except TV were used in the CCA. It was deliberately excluded as the time was too limited to process files from this source. In chapter 5, figure 11 has demonstrated the number of texts derived from each source and that most of the texts were obtained from magazines and then from websites. This is mainly because they are rich in a number of text types that are not available in newspapers. For example, texts such as short stories, autobiography, education, and tourism. In addition, it was easy to contact owners of magazines and get permission from them. It is as well easy to contact editors of newspapers but collecting texts from newspapers was delayed because the existing corpora are rich in this type of text.

Although the use of sources does not precisely match the selection of language teachers and engineers, this was to ensure a wide range of text types, with a reasonable number of words, were selected. Figure 23 shows that all sources that were equally important to both of them.

Figure 23: preferences of sources among language teachers and engineers



7.2 Contribution

Despite the growing awareness of the importance of corpora in the Arab community, the scale of this research is still in the early stages. The English-Arabic parallel corpus developed in Kuwait is the only one known to be used in teaching. It consists of 3M words and thus it is considered to be a prototype for a larger one. It is hoped that ten times this size would be

developed in the future to ‘meet the needs of researcher, translators, and English teachers in the Arab world.’ However, it is not certain if it is going to be free or users have to pay a fee. This project is the first attempt to promote the idea of a free corpus. Although it is still small and not all the text types have been included, texts that are ready have been made available and hopefully work on broadening the CCA will be continued.

In addition, this research has presented a thorough survey of the all the different Arabic corpora which are ready and which are under development. When this research started there was no idea about how many and what types of corpora were available. The number discovered was surprising. Nevertheless, this survey is far from being exhaustive as there are still other corpora for which information is difficult to find. For example, there is one in the King Abdulaziz City for Science and Technology, Saudi Arabia and also one by the Sakhr Company. However, such comprehensive survey might be useful for researchers who want to explore a particular linguistic aspect and who might want to get in touch with those owners of the corpora. As has been made evident in the previous chapters, all these corpora are either not freely accessible, or not provided with annotation. Having a free access to an Arabic corpus is important for learners, researchers, educators, and language engineers. There is a great gap in the Arabic research especially that which is built on real data.

There is also a need for foreign learners to be more familiarised with real Arabic rather than some artificial fragments in textbooks. The way to achieve this is by providing a well-designed free corpus which is consistent and coherent with the specific needs of the users. The teaching of Arabic faces numerous challenges based on the changing needs of learners and the use of technology in the domain of teaching languages. The recent research stresses the importance of incorporating communicative competency and cultural proficiency. In response to this new approach of teaching, many changes and improvements have been made in the development of materials, curriculum, and resources and of course in the training of teachers. Therefore, it is hoped that the CCA would contribute in this aspect and would be a valuable resource that matches the new philosophy of teaching Arabic. Until now there is a debate over what form of Arabic should be taught to foreigners. Although some teachers are more inclined to teaching MSA, this does not give the best solution if the aim to promote the skill of communication.

A general conclusion which can be derived from this research is that the road to having an Arabic corpus is not at all straightforward. Two problems have been encountered: one, lack of free taggers for Arabic. Two: unavailability of concordancers that work directly on Arabic texts without problems. Since written Arabic lacks vowels, it makes it difficult to use with

concordancers. Searching for key words is not always as accurate as it should be which is due to ambiguity.

A tagged corpus is much more useful for teaching and for conducting research. It is believed that Sakhr Company owns a tagged corpus but not willing to make it available for public use. Therefore, it is hoped that the CCA would be a resource to be used for testing newly developed taggers. Most of the Arabic corpora are created in Arabic windows. In order for one to use it he must have this system. The CCA is marked-up with XML and encoded in UTF-8, and therefore, to use it, systems are required to support these standards. It is hoped that this research and its findings will lead to the development of new processing tools that work better on Arabic and in developing new systems that work with Unicode and XML.

7.3 Future Development of the Corpus

For this project to be fully complete and to match the aims set at the outset, it needs more time and requires cooperation between the linguists and experts in information technology. There is a need for recruiting some professional Arab typists to handle transcribing the spoken recordings. The corpus still needs more detailed annotation, as well as proper part of speech tagging has to be carried out. Therefore, there is a need for an annotator who can develop a program that deals with annotation automatically. The collection and annotation of the texts had been handled manually, but if there is an automatic method that obtains material from the Web, it would be better for increasing the size of the corpus.

Since the corpus developed does not reflect the representativeness criterion of covering all the possible genres, it is hoped to find more resources that cover all the branches of knowledge. In addition, more sources need to be investigated and contacted for getting spoken recordings. Subsequently the material is expected to produce a valuable source relevant for not only wide ranging linguistics and language purposes, but also for language engineering applications.

An effort has been made to secure rights in order to include material in the corpus, but there is a need for a consortium of commercial members to increase its size more rapidly. Therefore, a plan has been made to extend the project and seek some outside members within the UK and in the Arab world for cooperation. A copy of the proposal is in appendix VI.

References

- Ahmad, K. and Corbett, G.** (1987). The Melbourne-Surry corpus. *ICAME* 11:39-43.
- Al-Ajmi, H.** (2002). Which microstructural features of bilingual dictionaries affect users' look-up performance? *International Journal of Lexicography*, Vol. 15 No.2. Oxford University Press. http://www3.oup.co.uk/lexico/hdb/Volume_15/Issue_02/.
- Al-Ajmi, H.** (2003). Compiling an English-Arabic parallel text corpus. In *Proceedings of Asian Association for Lexicography*, August 27-29, Meikai University, pp.51-54. <http://www.asialex.org/>.
- Al-Batal, M.** (1995). Issues in the teaching of the productive skills in Arabic. In *The Teaching of Arabic as a Foreign language*, Al-batal, M.,ed., Provo, UT: American Association of Teachers of Arabic, pp. 115- 133.
- Al-Muhanna, A.** (2003). *Scientific and technological terms transfer into Arabic: A corpus-based study of Arabic noun+noun and noun+adjective compounds*, Ph. D. thesis, UMIST.
- Alosh, M.** (1995). Computer-assisted language learning for Arabic: rationale and research potential. In *The Teaching of Arabic as a foreign language*, Al-batal, M., ed., Provo, UT: American Association of Teachers of Arabic, pp. 257-290.
- ANC** (2003). American National Corpus homepage. <http://americannationalcorpus.org/>.
- Atwell, E. et al.** (2000). User-Guided System Development in Interactive Spoken Language Education. *Natural Language Engineering Journal*, vol. 6 no.3-4, Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, pp. 229-241. [Online]. Available from World Wide Web: <http://www.comp.leeds.ac.uk/eric/nlejjournal2000.ps>.
- Belnap, K.** (1995). The institutional setting of Arabic language teaching: A survey of program coordinators and teachers of Arabic in U.S. institutions of Higher learning: In *The Teaching of Arabic as a Foreign language*, Al-batal, M., ed., Provo, UT: American Association of Teachers of Arabic, pp. 35- 77
- Bernard, L. and Dodd, T.** (2003). Xara: an XML aware tool for corpus searching. In *Proceedings of the Corpus Linguistics 2003 conference*. Volume 16, part 1. 142-144.
- Bernard, L.** (2004). BNC-Baby and Xaira. In *Proceedings of the Sixth Teaching and Language Corpora conference*. Granada, pp. 84.
- Boualem, M., Leisher, M. and Ogden, B.** (1996). *Concordancer for Arabic*. Computing Research Laboratory (CRL), New Mexico State University. [Online]. Available from World Wide Web: <http://crl.nmsu.edu/~mleisher/concord.pdf>.
- Elgibali, A. and Taha, Z.** (1995). Teaching Arabic as a foreign language: Challenges of the nineties: In *The Teaching of Arabic as a Foreign language*, Al-batal, M., ed., Provo, UT: American Association of Teachers of Arabic, pp. 79- 102

- Baker, P. et al.** (2003). Constructing corpora of South Asian languages. In *Proceedings of the Corpus Linguistics 2003 conference*. Volume 16, part 1. 71-80.
- Beesley, K.** (2001). *Finite-State morphological analysis and generation of Arabic at Xerox research: status and plans in 2001*. In <http://www.elsnet.org/acl2001-arabic.html>.
- Beesley, K.** (2003). *Xerox Arabic Morphological Analyser Surface-Language* [Online]. Available from World Wide Web: ([Unicode](http://www.unicode.org/unicode/Documentation/)) [Documentation](http://www.xrce.xerox.com/competencies/content-analysis/arabic-inxight/arabic-surf-lang-unicode.pdf).
<http://www.xrce.xerox.com/competencies/content-analysis/arabic-inxight/arabic-surf-lang-unicode.pdf>.
- Bell, J. and Zemanek, P.** (1994). *Test of two Arabic OCR programs* [Online]. Available from World Wide Web: www.hf.uib.no/smi/ksv/arabocr.html.
- Berri, J., Zidoum, H. and Atif, Y.** (2001). Web-based Arabic Morphological analyser. In *CICLing 2001*, Gelbukh, A ed, LNCS 2004, pp. 216-225, 2001. Springer-Verlag Berlin Heidelberg.
- Biber, D., Conrad, S. and Reppen, R.** (1998). *Corpus linguistics. Investigating language structure and use*. Cambridge University Press.
- British National Corpus.** <http://www.natcorp.ox.ac.uk/>.
- Brown, J.** (2001). *Using surveys in language programs*. Cambridge University Press.
- Buckwalter, T.** (2002). Xerox's Corpus [Online]. Available from the World Wide Web: <http://www.qamus.org/wordlist.htm>.
- Cobb, T.** (1997). Is there any measurable learning from hands-on concordancing? *System*, 25, 301-315.
- Cobb, T., Greaves, C. and Horst, M.** (2001). Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In *Regards sur la didactique des langues secondes*, Raymond, P. and Cornaire, C., Montréal: Éditions logique.
- Darwish, K.** (2002). *Building a shallow morphological analyser in one day*. <http://www.cs.umd.edu/Library/TRs/CS-TR-4326/CS-TR-4326.pdf>.
- Dodd, B.** (1997). Exploiting a corpus of written German for advanced language learning. In *Teaching and language corpora*, Wichmann, A. et al., eds., Longman, pp. 131-145.
- Elkhafaifi, H.** (2001). Teaching listening in the Arabic classroom: a survey of current practice. *Al-^cArabiyya* 34, pp. 55-90.
- Elliott, D., Hartley, A. and Atwell, E.** (2003). Rationale for a multilingual corpus for machine translation evaluation. In *Proceedings of the Corpus Linguistics 2003 conference*, University Centre for Computer Corpus Research on Language. Vol 16 part 1, 191-200.
- ELRA** (2003). European Language Resources Association homepage. <http://www.elra.info/>.
- Fang, A. C.** (1992). Building a corpus of computer science English, In *English Language Corpora: Design, Analysis and Exploitation*, Aarts et al, eds., Amsterdam: Rodopi. pp 73-78.

- Ferguson, C. A.** (1963). Problems of teaching languages with Diglossia. *Georgetown University Monograph Series on Languages and Linguistics*. 15: 163-177.
- Freeman, A.** (2001). *Brill's POS tagger and a morphology parser for Arabic* [Online]. Available from the World Wide Web: <http://www.elsnet.org/acl2001-arabic.html>.
- Freeman, A.** (2002). What is a word? In *Proceedings of the International Symposium on: The Processing of Arabic*, Tunisia, pp. 31-44.
- Graddol, D.** (1997). *The future of English*. London: British Council.
- Haddad, S.** (1985). Tadrīs al-mahaaraat al-shafawiyya: mawqif jadiid. *Al-ʿArabiyya* 18 (1 & 2): 15-21.
- Hadley, G.** (2001). *Concordancing in Japanese TEFL: Unlocking the Power of Data-Driven Learning*. [Online]. Available from the World Wide Web: <http://www.nuis.ac.jp/~hadley/publication/jlearner/jlearner.htm>
- Haslerud, V. and Stenström, A.** (1995). The Bergen corpus of London teenager language (COLT). In *Spoken English on computer*, Leech, G et al,eds., Longman, pp. 235-242.
- Holes, C.** (1990). A Multi-media, topic-based approach to university-level Arabic language teaching. In *Diglossic Tension: teaching Arabic for communication*, Aguis, D.,ed., Beaconsfield Papers. Leeds: Folia Scholastica, pp. 36-41.
- Hoogland, J.** (1996). The use of OCR software for Arabic in order to create a text corpus of Modern Standard Arabic for lexicographic purposes. *Proceedings of the international conference and exhibition on multi-lingual computing*, pp.2.7.1-2.7.16.
- Hoogland, J.** (2003). http://www.let.kun.nl/wba/Content2/1.4.6_Concordancing.htm.
- Ide, N.** (2003). The American National Corpus: Everything you always wanted to know and weren't afraid to ask. Invited keynote, Corpus Linguistics 2003, Lancaster, UK. (ppt presentation).
- Izwaini, S.** (2003). Building specialised corpora for translation studies. In *Proceedings of the workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, Lancaster 27 March 2003, pp.17-25.
- Johansson et al.** (1986). The tagged LOB Corpus. Users' manual. The Norwegian Centre for the Humanities, Bergen
- Johns, T.** (1988). Whence and whither classroom concordancing? In *Computer applications in language learning*, Bongaerts, T., et al.,eds. Dordrecht: Foris.
- Jones, R.** (1997). Creating and using a corpus of spoken German. In *Teaching and language corpora*, Wichmann, A. et al, eds, Longman, pp. 67-82.
- Kanungo, T., Marton, G., and Bulbul, O.** (1998). *Performance Evaluation of two Arabic OCR Products*. [Online] <http://www.cfar.umd.edu/~kanungo/pubs/aipr.pdf>.
- Khoja, S.** (2001). APT: Arabic part-of-speech tagger. In *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*, Carnegie Mellon University, Pittsburgh, Pennsylvania.

- Khoja, S.** (2003). *APT: An automatic Arabic part-of-speech tagger*. PhD thesis. University of Lancaster.
- Khoja, S. et al.** (2003). A tagset for the morphosyntactic tagging for Arabic. In *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Wilson, A. et al, eds., Lincom-Europa, Munich, pp.59-72.
- Kučera, H. and Francis, W.H.** (1967). *Computational analysis of present-day American English*. Brown Univeristy Press, Providence, Rhode Island.
- Leech, G.** (1993). 100 million words of English: the British National Corpus (BNC) Project. English Today.
- Leech, G.** (1997). Teaching and Language Corpora: a Convergence. In *Teaching and language corpora*, Wichmann, A. et al,eds., Longman, pp. 1-23.
- Leech, G. and Wilson, A.** (1999) Standards for tagsets. In *Syntactic Wordclass Tagging*, Van Halteren, H., ed., Kluwer, Dordrecht, pp. 55-80.
- Maamouri, M. and Cieri, C.** (2002). Resources for Arabic Natural Language Processing at the linguistic Data Consortium, In *Proceedings of the International Symposium on: The Processing of Arabic*, Tunisia , pp.125-146.
- Meyer, C.** (2002). *English corpus linguistics, an introduction*. Cambridge University Press.
- Minugh, D.** (1997). All the language that's Fit to Print: using British and American Newspaper CD-ROMs as Corpora. In *Teaching and language corpora*, Wichmann, A. et al,eds., Longman, pp. 67-82.
- van Mol, M.** (2000). The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach. In *Proceedings of the ninth EURALEX International Congress*, Stuttgart, 8-12 August, pp. 831-836. [Online] http://www.ilt.kuleuven.ac.be/ilt/arabic/_pdf/stuttgart.pdf.
- Nikkhou, M. and Choukri, K.** (2004). *Report on survey on Arabic language resources and tools in the Mediterranean countries*. [Online]. Available from the World Wide Web: http://www.nemlar.org/Publications/NEMLAR-REPORT-SURVEY-FINAL_web.pdf.
- Parkinson, D.** (1995). Computers and proficiency goals. In *The Teaching of Arabic as a Foreign language*, Al-batal M., ed., Provo, UT: American Association of Teachers of Arabic, pp. 291-317.
- Peters, P.H.** (1987). Towards a corpus of Australian English. *ICAME* 11: 27-38.
- Roberts, A.** (2004). *aConCorde*. <http://www.comp.leeds.ac.uk/andyr/software/aConCorde/>
- Sharoff, S.** (2004). Towards basic categories for describing properties of texts in a corpus. In *Proc. of Language Resources and Evaluation Conference (LREC04)*. May, 2004, Vol. V, Lisbon, Portugal, pp.1743-1746.
- Shastri, S.V.** (1988). The Kolhapor corpus of Indian English and work done on its bases so far. *ICAME* 12: 15-26.
- Sinclair, J.** (1996). Preliminary recommendations on text typology. Eagles Document EAG-TCWG-TTYP/P. <http://www.ilc.cnr.it/EAGLES96/texttype/texttyp.html>.

- Stevens, V.** (1991). Concordance-based vocabulary exercises: A viable alternative to gap-fillers. In Johns, T. and King, P., eds. *Classroom concordancing: English Language Research Journal*, 4: 47-63. University of Birmingham: Centre for English Language Studies.
- Stubbs, M.** (1996). *Text and Corpus Analysis*, Oxford: Oxford University Press.
- Svartvik, J.** (1990). The London-Lund corpus of spoken English: description and research. Lund University Press, Lund, Sweden.
- Taha, Z.** (1995). The grammar controversy: what to teach and why? In *The teaching of Arabic as a foreign language*, Al-batal M., ed., Provo, UT: American Association of Teachers of Arabic, pp, 175-183
- Taylor, L. J. and Knowles, G.** (1988). Manual of information to accompany the SEC sorpuc: the machine readable corpus of spoken English. University of Lancaster.
- Thomson, W. M.** (1994). The teaching of Arabic in universities: A question of balance. Leeds Arabic Papers, Department of Modern Arabic Studies, University of Leeds.
- Tiomajou, D.** (1993). Designing a corpus of Cameroonian English. *ICAME* 17:119-124.
- Voshmgir, S.** (1999). XML Tutorial (Java Commerce.com). <http://www.javacommerce.com/tutorial/xmlj/intro.htm>.
- Younes, M.** (1990). An integrated approach to teaching Arabic as a foreign language. *Al-^cArabiyya* 23 (1&2): 105-122.
- Zemanek, P.** (2001). *Clara (Corpus Linguae Arabicae): An Overview*. [Online]. Available from the World Wide Web: <http://www.elsnet.org/acl2001-arabic.html>.

Appendix I Samples of Existing Corpora

Sample 1

Buckwalter Corpus:

مجلة نيتشر الدورية

كشف النقاب عن ملاذ لفيروس أتش أي في >--

العلماء يأملون في تعقب كافة الخلايا المصابة بفيروس أتش أي في لدى المرضى تقول دراسة علمية حديثة نشرتها دورية نيتشر العلمية إن العلماء قد نجحوا في اكتشاف الكيفية التي يتمكن بها فيروس نقص المناعة المكتسب (أتش أي في) من التأثير على مفعول أحدث العقاقير المقاومة للفيروس مما يعوق استئصاله. ويمكن لهذا الاكتشاف أن يسهم في النهاية في المزيد من العلاجات الفعالة. فبينما يمكن للعقاقير الحديثة أن تتخلص من فيروس أتش أي في فإن القليل منها يتمكن من تجنب الدمار وتشكيل "خزان" في مكان ما بالجسم. وعندما يتوقف مفعول العقار المضاد للفيروسات فإن الفيروس يظهر ثانية. وهذا يعني بدوره أن على المرضى أن يتعاطوا عقاقير باهظة الثمن لمدى الحياة من أجل السيطرة على الفيروس، وليس هناك ضمان على أن فعالية العقاقير لن تتضاءل بمرور الوقت. وتعقب الخزان بحيث يمكن اكتشاف طريقة لمهاجمته بشكل أولوية بالنسبة لباحثين في كلية ماساتشوستس الطبية. وفي الظروف العادية، يهاجم فيروس الإيدز خلايا الجهاز المناعي المسماة بـ"خلايا تي". وساد الاعتقاد بأن فيروس أتش أي في يمكن فقط أن يصيب هذه الخلايا عندما تكون نشطة، بحيث أنها تستجيب لـ"الغزاة". ولكن الأدلة توحي بأن الفيروس وجد طريقة غير مباشرة لإصابة خلايا تي حتى عندما تكون في حالتها غير النشطة. علاجات أفضل ولن يمكن الوصول إلى فيروس أتش أي في المختبيء في هذه الخلايا بواسطة العلاجات الحالية المضادة للفيروس.

الهدف لن يقتصر على تقليل إنتاج فيروس أتش أي في-1 ولكنه سيتطرق أيضا إلى تدمير خزانات الخلايا التي تحول حتى الآن دون القدرة على استئصال عدوى الإصابة بفيروس أتش أي في-1

روجر بوميرانتر الباحث في جامعة توماس جيفرسون المختص في حقل فيروس أتش أي في واكتشف فريق البحث أن فيروس أتش أي في كان قادرا على إصابة خلية منفصلة تعرف بالخلية الأكلة الكبيرة. وهذا بدوره يحثها على إفراز مادة كيميائية تعدل من سلوك خلية تالفة تدعى "خلية بي". وعندما تتصل خلايا بي المعدلة بخلايا تي غير النشطة أو خلايا تي الخاملة فإنها تجعلها عرضة للعدوى بالفيروس. وقال روجر بوميرانتر الباحث في جامعة توماس جيفرسون المختص في حقل فيروس أتش أي في إن اكتشاف هذا التفاعل المعقد يمكن أن يساعد العلماء في أبحاثهم الرامية إلى اكتشاف طرق لاستئصال الفيروس. ونظريا، ربما يكون من الممكن اعتراض أو منع هذه العملية عن طريق منع الفيروس من العثور على ملاذ في خلايا تي غير النشطة. وقال بوميرانتر: "الهدف لن يقتصر على تقليل إنتاج فيروس أتش أي في-1 ولكنه سيتطرق أيضا إلى تدمير خزانات الخلايا التي تحول حتى الآن دون القدرة على استئصال عدوى الإصابة بفيروس أتش أي في-1."

#F bbc_031104_sci_tech/newsid_3082000/3082294.stm

آخر تحديث: الخميس 04 سبتمبر 2003 GMT19:13

إلقاء القبض على مشتبه به آخر في فيروس MSBlast

فيروسات الكمبيوتر الفيروسات الإلكترونية تزداد شراسة في عام 2003
الدودة تهاجم الحواسيب المنزلية
"سوبيج" يعيث فسادا في أجهزة الكمبيوتر
اتهام رسمي لشاب أمريكي بشأن فيروس إنترنت
العالم يستيقظ على فيروس جديد

فيروس كمبيوتر جديد يتسلل كرسالة ادارية
شركة بيت ديفيندر
جامعة لاسي الرومانية

Sample 2

Leuven Corpus

السيد عيد العربي نُرحب بكم بدايةً . شعار مؤتمركم هو خطاب الحقيقة، إجماع وطني، الارتباط بالقيم الإنسانية والثقافية العالمية . لماذا استعملتم عبارة القيم الإنسانية وليس القيم الوطنية وهل يعني هذا أن القيم الوطنية تستثني القيم الإنسانية .

شكرا أولاً باسم قيادة التجمع من أجل الثقافة والديمقراطية أقدم شكرنا للإذاعة الوطنية القناة الأولى على هذه الدعوة . وبودّي أن أقدم تصحيحاً هو أنّ هذا المؤتمر هو المؤتمر الأول وليس المؤتمر التأسيسي لأنّ المؤتمر التأسيسي قد عُقد او انعقد في ففري تسعة وثمانين . وهذا المؤتمر صدر عن برنامج وقيادة وطنية وكذلك فنقول المؤتمر الأول للتجمع . وبالنسبة للسؤال، خطاب الحقيقة نلاحظ أنّه منذ أكثر من ربع قرن وشعبنا لم يطلع عما يجري في البلاد كأنه متفوّج وليس صاحب السيادة ليقرّر وفقاً لمعطيات وحقائق وبعد ذلك يتخذ ما يظن أنه لأخيراً هو صاحب السيادة .

بالنسبة للإجماع الوطني : هناك مشاكل كثيرة تعانيها البلاد ولا يمكن لأيّ حزب أو تشكيلة سياسية أن تقوم بحلّ هذه المشاكل أو بإخراج المجتمع الجزائري منها ولذلك فنحن ننادي بإجماع وطني لأنّ القضية انتاعهم، كلّ الديمقراطيين لندفع بالمسار الديمقراطي إلى نقطة اللار جوع . أما بالنسبة لارتباطنا بالقيم الإنسانية في، والثقافية في العالم هذا لا يعني أننا نتخلّى عن القيم الوطنية . القيم الوطنية بالدرجة الأولى، ولكننا مقبلين على الألفية الثالثة فنتفحناً على العالم ضروري إن لم نريد نكون من المنسيين .

سي عيد العربي، المؤتمر إنّه ينعقد غداً وهو المؤتمر الأول كما تفضلتم باختصار تقييمكم . لأهمّ المراحل السياسية والتنظيمية التي قطع حزبكم، حزبكم منذ نشأته؟

نعم . أظنّ أنّه من الضروري أن نذكر بأنّ مناضلي التجمع من أجل الثقافة والديمقراطية كانوا مناضلين للحقوق الثقافية وللديمقراطية ولحقوق الإنسان بصفة عامة منذ سنوات وسنوات ولقد نادينا بالديمقراطية وبالحقوق الثقافية قبل إنشاء التجمع . ولكن بعد حوادث أكتوبر ترتأينا أن ننشئ تنظيمًا سياسيًا ليأخذ كلّ هذه الجوانب في برنامج سياسي يُقدّم للشعب الجزائري ليقول فيه كلمة وهو برنامج من جملة البرامج التي تُطرحها الأحزاب السياسية بعدما سمح دستور ثلاثة وعشرين فيفري بالتعددية السياسية .

Sample 3

(LDC)

a. The Written Corpus:

<DOC docid="ANN20021215.0019">

<h1>

<seg id=1> بعدما أصبحا شريكين واستحالت الحكومة وقوداً لعلاقتهما
</seg>

</h1>

<p>

<seg id=2> كتب نقولا ناصيف </seg>

</p>

<p>

<seg id=3> لم يكن رئيس الحزب التقدمي الاشتراكي النائب وليد جنبلاط ولا الوزير طلال ارسلان في حاجة الى اتفاق الهاتف الخليوي كي يثيرا في وجه الرئيسين اميل حود ورفيق الحريري عاصفة رد الاعتبار الى مجلس الوزراء </seg>

<seg id=4> اذ منذ تأليف حكومة ما بعد انتخابات 2000 ومجلس الوزراء يفقد تدريجاً دوره تارة في ظل تفاهم الرئيسين (في المرحلة الاولى من عمر هذه الحكومة) وطوراً في ظل نزاعاتهما المفتوحة الى ان أوقفت </seg>

<seg id=5> ولم ينجح مجلس الوزراء على امتداد السنتين الاخيرتين في الاضطلاع بدور صمام الامان في العلاقات المترجحة بين الرئيسين الى حد أضحى </seg>

<seg id=6> وعلى مر هاتين السنتين أثير أكثر من مرة مرسوم تنظيم </seg> أعمال مجلس الوزراء وضرورة التزام احكامه كاملة، على ان معيار الاستقرار الحكومي واضطرابه استمر رهن العلاقة الشخصية بين حود والحريري اللذين لم يتحمسا لاقتراح يجعل مرسوم تنظيم أعمال مجلس الوزراء قانوناً </seg>

</p>

<p>

<seg id=7> وفي الايام الاخيرة أعيد طرح هذا الموضوع، وبدا المقصود منه توجيه اكثر من انتقاد الى رئيس الجمهورية ورئيس الوزراء على السواء </seg>

<seg id=8> ومع ان ثمة من يعتقد بأن تبني جنبلاط وارسلان المطالبة </seg> باحياء مرسوم تنظيم أعمال مجلس الوزراء يخرج رئيس الجمهورية ولكنه يستهدف رئيس الحكومة، فان اكثر من وجهة نظر ترى مسؤوليتهما </seg>: متساوية في تعطيل دور مجلس الوزراء لاسباب شتى منها </p>

</p>

<p>

<seg id=9> ان رئيس الجمهورية ورئيس الوزراء قد اتفقا منذ اكثر 1- </seg> من سنة ونصف اثر آخر تصويت في مجلس الوزراء على مشروع "سوليدير"، على "تنظيم" ادارتهما لاعمال مجلس الوزراء في معزل عنه، وهذا ما عبّر عنه تفاهمهما على عدم ادراج اي بند في جدول الاعمال لا يكونان قد وافقا عليه </seg>

<seg id=10> وتالياً، فانهما بتصرفهما هذا يكونان قد اختصرا في </seg> دوريهما مجلس الوزراء، فلا يُطرح عليه الا المشروع الذي يحظى باتفاق الرئيسين، على ان يعني ذلك في المقابل ان طرح المشروع في جلسة مجلس الوزراء ينبغي ان يقود الى اقراره على النحو الذي رعاه اتفاق الرئيسين </seg>

</p>

<p>

UMAAH_UM.ARB_backissue_39-a.0016

إعادة انتخاب البشير رئيساً للحزب الحاكم-53

جدد حزب المؤتمر الوطني الحاكم في السودان ولاية الرئيس عمر البشير رئيساً للحزب لدورة -54 جديدة تستمر عامين، وبدأ مناقشة قضايا مثيرة للجدل تتصل بإقرار السلام والمصالحة الوطنية وإجراء تعديلات دستورية.

وتعهد البشير، عقب انتخابه في المؤتمر العام للحزب الذي يحضره نحو ستة آلاف عضو ووفود عربية وأفريقية، بأن يظل حزبه في طليعة القوى السياسية في البلاد مهما تبدلت الخارطة السياسية

وكرر رفض بلاده الحملة العسكرية الأمريكية على أفغانستان لأنها تناول الأبرياء، وطالب بتعريف دولي للإرهاب.

وأكد الأمين العام للحزب الدكتور إبراهيم أحمد عمر أن حكومته لن تتراجع عن الشريعة الإسلامية ومبادئها، فيما دعا مسجل الأحزاب محمد أحمد سالم الحكومة إلى دعم الأحزاب مالياً من خزنة الدولة. بمعايير يحكمها ثقل الأحزاب في البرلمان حتى يكون دعماً للديموقراطية.

وبدأ المؤتمر الذي يعقد للمرة الأولى في اجتماع عادي منذ إنشاء زعيم الحزب السابق الدكتور حسن الترابي حزباً مستقلاً العام الماضي، في مناقشة أوراق عدة أبرزها الورقة السياسية التي دعت إلى إجراء تعديلات دستورية تشمل استحداث منصب وزير أول (رئيس وزراء) الذي لم يستخدم منذ تقلد الرئيس عمر البشير السلطة في العام 1989، وإقامة مجلس للشيوخ وتوسيع سلطات الرئيس ومنحه صلاحيات إعفاء حكام الولايات والدعوة إلى انتخابات برلمانية مبكرة.

لكن مسؤولين في الحكومة يطالبون بإجراء إصلاحات ديموقراطية داخل الحزب وتعديلات دستورية تسمح بانتخاب حكام الولايات انتخاباً مباشراً وهي القضية التي أدت إلى حل البرلمان السابق الذي كان يرأسه الترابي.

وأكد وزير الزراعة مجذوب الخليفة والنقل الدكتور لام اكول في تصريحات للصحافيين يوم 10/17 الحالي تمسكهما باختيار حكام الولايات عبر انتخابات مباشرة.

18/10/2001 الحياة ص 10-55

ار 0223 4 ش 0064 قبر /افب-غمح 06 اسر انيل/فلسطينيون 20001115_AFP_ARB.0220
وفاة فلسطيني متأثراً بجروحه في خان يونس

غزة 11-15 (اف ب)- افاد مصدر طبي فلسطيني اليوم الاربعاء ان فلسطينيا توفي متأثرا بجروح اصيب بها ظهر اليوم خلال مواجهات وقعت بين الجيش الاسرائيلي وشبان فلسطينيين. وافاد المصدر الطبي ان الطفل جهاد ابو شحمه (12 عاما) الذي اصيب بعيار ناري في رأسه قد توفي متأثرا بجروحه التي اصيب بها اليوم في خان يونس.

افب صخر/س ع ها ———

b. The Spoken Corpus (News Broadcast):

45.44 47.99 : <English The following program is in Arabic>

56.58 60.75 : hunA <N wA\$untun> mustamicInA fl \$amAli (Ca)afrIqiyA (Ca)ascada
Allahu SabAHakum

60.78 64.66 : Al (Ci)ivAcaBu Al carabiy*aB li Sawti (Ca)amirIkA tuHayyIkum wa
tuqadimu lakum

64.65 68.66 : havihi Al fatraBa Al (Ci)ixbAriyyaB calE Al mawjAti Al qiSAr xamsaB wa
ci\$riIn

69.33 74.47 : wa xamsaB wa ci\$riIn fASil (Ca)arbacaB wa (th)alA(th)In fASil wAHid

74.96 77.10 : wa wAHid wa (th)alA(th)In fASil (Ca)arbacaB mitr

77.21 82.96 : wa bi Al <B kIU hirtz> (Ci)iHdA ca\$ara (Ca)alf wa tiscA(CI)aB wa
xamsaB wa tiscIn (Ci)iHdA ca\$ara (Ca)alf wa (th)amAnmA(CI)aB wa sitIn

83.19 87.73 : (Ci)iHda ca\$ara (Ca)alf wa (th)amAnmA(CI)aB wa xamsaB tiscaB
(CA)alAf wa subcmA(CI)aB wa xamsaB ca\$ar

87.95 90.73 : wa tiscaB (CA)alAf wa sitmA(CI)aB wa sitIn

105.44 107.75 : wa havihi (Ca)awala(an) mustamicInA fl \$amAli

107.74 111.10 : (Ca)afriqiyA canAwInu na\$raBi (Ca)anbACi havA Al SabAH

111.19 116.02 : Al wilAyAtu Al mut*aHidaB tatawaqacu (Ci)ijrACa muHAda(th)Ati Al
salAm fl Al \$arqi Al (Ca)awsaT

116.62 121.97 : fl Al mawcid.. fl Al mawcidi Al muqarar raGma ma\$Akili HukUmaBi
ra(CI)Isi Al wuzarAC <N (Ci)IhUd bArAk>

123.01 125.66 : sUryA tanfl (Ca)an*a wazIra Al xArijiy*aB fArUq Al \$arc

125.67 129.86 : qAl (Ci)in*a (Ci)insiHAbA Al jay\$A Al sUrI min lubnAn sa yuCowadI
(Ci)ilE Harb(in) (Ca)ahliy*aB

131.47 135.29 : wazIru Al xArijiy*aB Al sUdanI muSTafE cu(th)mAn (Ci)ismAcII
yaqUI

135.32 139.87 : (Ci)in*a HiwArA(an) (Ci)IjAbiy*A(an) yadUru Al (CA)an bayna
bilAdihi wa Al wilAyAt Al mut*aHidaB

c. The Conversational Corpus: CALLFRIEND:

375.33 375.85 B: %ah

375.79 377.51 A: yaeni \$aGGAl nabaT\$iyitEn wara bacD kull yOm dilwaqti

377.47 379.10 B: [static/] bitAxud nabaT\$iyitEn wara bacD [/static]

378.92 379.35 A: A

379.74 380.02 B: %ah

380.18 380.65 A: bass

381.19 381.57 B: %ah

381.62 383.34 A: il+muhimm kallimtili il+nAs dOI bi+il+amAnaB~

383.68 384.18 B: A

384.62 385.44 B: &samyaB~ ahi macAk

385.74 386.47 A: baqullak Eh

386.78 387.23 B: %ah

387.12 389.11 A: wallAhi il+caZIm kallimtuhum walla bitqulli kida bass

388.60 389.96 B: aywa ya &sA- aywa ya &sAmi

390.02 391.27 A: bit- bithaddIni yacni

391.53 392.44 B: la la la IEh

392.51 394.09 A: yacni Hazcal wallAhi bi+gadd (())

393.78 394.44 B: la la la la

394.58 395.69 A: il+mawDUc da muhimm qawi aSlu

396.01 396.70 B: la OkkE

397.16 397.47 A: %ah

397.76 398.56 B: &samyaB~ ((ahi wayyAk ahi))

400.92 401.71 B1: %ha ya &simsim

Sample 4

Nijmegen Arabic Corpus

file: bioar003.doc

عام 18 م عرض رئيس الولايات المتحدة الأمريكية على زعيم قبيلة الدواميش الهندية إسياتل < أن يبيع أرضه ، وأن ينتقل مع شعبه الى إحدى المحميات ، فأجاب الزعيم سياتل : <أنى للإنسان ان يبيع أو يشتري أجزاء من الأرض ؟ وكيف (يتملك) الإنسان ما لا يخصه ؟ ربما يسعفنا تفكيرنا في تفهم ذلك لو علمنا بماذا يحلم الإنسان الأبيض . . والتصورات التي يزرعها في أذهان أبنائه . >

وبعد ما يربو على < 13 سنة عاد الرجال البيض إلى خيام الهنود، لكنهم لا يرتدون الزي العسكري هذه المرة بل ملابس ملائكة الرحمة، ناصعة النياض ، ولا يرغبون في أراضي الهنود بل في شعرهم ، وجلودهم ، ودمائهم ! فمنذ او اخر ثمانينيات القرن العشرين بدأ العلماء البيض بزيارة هنود الغوايمي في باناما، ليأخذوا في البدء من فئة قليلة ثم من جميع افراد القبيلة، عينات من الدم ، وارسلوها إلى المختبرات الأمريكية. فماذا كانت النتيجة؟

<<صرعة> طبية منقطعة النظير ! إذ اكتشف الدكتور الأمريكي مايكل ليرمور تماثل سيدة من الغوايمي (26 سنة) للشفاء من الداء العضال : ابيضاض الدم (اللوكيميا) من تلقاء ذاتها، ودون عون طبي أو دوائي! إذ احتوى دمها على مضادات للفيروس المحرض لابيضاض الدم ، المشابه من حيث البنية الأساسية

للفيروس المحرض لمتلازمة نقص المناعة البشرية المكتسب : الإيدز. وليس غريباً أن يثير هذا <الفتح> المذهل في الوسط العلمي العديد من التساؤلات مثل:
 1- هل يسمح حل <لغز> الشفرة الوراثية لهذه المضادات الواقية باختكار طرق سريعة وفعالة لتشخيص المرض؟
 2- ألا تساعد الشفرة الوراثية (الجينية) هذه على إنتاج اللقاحات التي تمنع الإصابة اصلاً؟
 تضمن رد الزعيم سياتل بان <الإنسان لم يخلق نسيج الحياة، التي لا يشكل فيها-هو نفسه- أكثر من ليف وحسب >. وحسب اعتقادي المتواضع ما كان ليخطر على قلب هذا الزعيم - حتى في أشد حالاته تشاوما- أن يأتي حين من الدهر يدعي فيه شخص أبيض ان مورثات أحد الهنود تنضوي تحت لائحة >> اختراعاته. >

Sample 5

CLARA

ArLing_01

ووضع الخليل اضافة الى الحركات علامات للهمزة والتشديد والروم والاشمام. " أولية النحو اختلفت الآراء قديماً وحديثاً فيمن وضع اللبنة الأولى في النحو العربي. فقد جاء في رواية ان الامام على بن ابي طالب هو واضع هذا العلم وذلك بسبب لحن سمعه أراد تقويمه ، فهداه تفكيره الى وضع أصول العربية.
 ويقال : ان ابا الاسود الدؤلي هو الذي وضع اصول العربية وذلك حين اضطرب كلام العرب فغلبت السليقية ولم تكن نحوية ، فكان سرارة الناس يلحنون ووجوه الناس ، فوضع باب الفاعل والمفعول به ، والمضاف ، وحروف الرفع والنصب ، والجر ، والجزم ، وقيل : ان عمل ابي الاسود كان باشارة من الامام على أو من زياد بن ابيه ، أو من عبيد الله بن زياد.
 وانكر رأي آخر على ابي الاسود هذا الصنيع ونسبه الى عيد الرحمن بن هرمز أو نصر بن عاصم الليثي ونسب الى ابي الاسود صنيع آخر غير وضع اصول العربية وهو نقط المصحف " أي وضع الشكل " لان المصاحف قبله - كما تقدم - كانت خالية من النقط مما ادى الى الخطأ في قراءة القران الكريم.
 ولقد انكر الاستاذ ابراهيم مصطفى من المحدثين هذه الاقوال وعدها حديث خرافة وفسر نسبة القدماء وضع النحو لابي الاسود بالتوهم والخلط وذلك أن القدماء خلطوا بين نقط المصحف ووضع النحو ، وسبب الخلط أن ضبط الكلمات كان يسمى نحواً ، فظن القدماء الذين جاءوا بعده أنه وضع النحو بالمعنى الاصطلاحي المعروف.
 وقد جعل الاستاذ مصطفى كتب النحو المعروفة حكماً في ذلك ، فوجد ان ايا منهم لم ينسب رأياً نحويّاً الى ابي الاسود فلو صحت الرواية التي تنسب وضع النحو اليه لوجدنا آراء له في الكتب النحوية المعروفة.
 وعلى مقولة الاستاذ ابراهيم مصطفى هذه يكون أول واضع للنحو هو اقدم نحوي تنسب اليه آراء نحوية.
 وقد وجد ان ابن ابي اسحق الحضرمي هو اقدم نحوي من هذا النوع.
 والفيصل في هذه المسألة هو وضع تعريف محدد للعمل النحوي ، فان كان المقصود بالنحو وضع اصول متطورة بعض التطور فلا شك ان ابن ابي اسحاق هو الرائد الاول في هذا الميدان لما له من ذكاء وجهود في هذا المضمار.
 اما اذا كان المقصود بالنحو مجرد وضع اصول اولية بدائية كرفع الفاعل ونصب المفعول فلا شك أن وضع النحو بهذا المعنى يرقى الى زمن سابق لابن ابي اسحاق اذ لا يعقل أن يكون ابن ابي اسحاق وحده هو الذي وضع النحو او الاصول

Sample 6:

Al-Hayat Corpus

1419-09-01 : هـت، 1998-12-19 : م.ت
14 : الصفحة، 13073 : العدد

أودي «أولرود كواترو» ستشق طريقها الى الإنتاج

أعلنت أودي في نهاية الأسبوع الماضي إتخاذ قرار نقل نموذج «أولرود كواترو» الى حيز الإنتاج إبتداء من شباط (فبراير) سنة 2000.

وسيستثمر الصانع الألماني (جزء من مجموعة فولكسفاغن) 105 ملايين مارك ألماني (نحو 636 ألف دولار أميركي) في مصنع نيكارسولم الذي سينتج الموديل الجديد على أساس موديل «أي 6 أفانت» ذي الصندوق الممدود (واغن).

وسيجهز «أولرود كواترو»، الذي عرض للمرة الأولى أوائل السنة الجارية في معرض ديترويت الدولي للسيارات، بنظام تعليق هوائي قابل لتعديل الإرتفاع في ثلاثة مستويات، إضافة الى علبة تحويل تسمح بإستغلال دفعه الرباعي بمجموعي نسب، الأولى للقيادة العادية والأخرى قصيرة للتعامل مع المسالك الشديدة الوعرة.

من ناحية المحرك سيتوافر جديد أودي بخياري محركين ذوي ست أسطوانات (V)، أولهما بنزني سعته 2.7 ليتر (30 صماما) ومدعوم بشاحنين توربينيين (معروف في فئة «إس 4»، وقوته فيها 265 حصانا وعزمه 400 نيوتون-متر)، والثاني توربو ديزل مع تقنية البخ المباشر، وسعته 2.5 ليتر (24 صماما، وقوته 150 حصانا).

وسيلحظ إختلاف الموديل الجديد بوضوح عن فئة «أي 6 أفانت» بفوارق شكلية تشمل المصابيح والمصددين والحماية السفلية وتصميم العجلات (16 أو 17 بوصة) وأقواسها المنفخة بعض الشيء (لإضفاء طابع خشونة نسبية تتجانس مع إمكانات الخروج عن الطرقات المعبدة)، والحافتين الجانبيتين السفليتين وشبك التهوية الأمامي.

وسيتوجه «أولرود كواترو» الى قطاع السيارات السياحية الفخمة والمؤهلة في الوقت ذاته للتعامل مع ظروف طبيعية صعبة.

ويشار الى أن مصنع نيكارسولم الذي سينتج الموديل الجديد، ينتج أيضا موديلي «أي 6» الكبير (في صيغتي الصالون بأربعة أبواب والواغن بصندوق ممدود) و«أي 8» الفخم العالي الذي يعتمد الألومنيوم في بنيته الداخلية وفي ألواح الخارجية، علما بأن المصنع ذاته سينتج أيضا أواخر السنة المقبلة موديلًا صغيرا (عرف حتى الآن في نموذج «أي إل 2») سيستخدم، مثل «أي 8»، الألومنيوم في بنيته الداخلية وفي ألواح الخارجية على حد سواء.

وكان مصنع نيكارسولم (يستخدم 11800 شخص) أنتج في الأشهر العشرة الأولى من السنة الجارية نحو 160 ألف سيارة، علما بأن أودي (تنتج في مصانع أخرى في ألمانيا وهنغاريا والصين وجنوب إفريقيا) باعت العام الماضي 546 ألف سيارة في العالم

الموضوع: سيارات

Sample 7:

An-Nahar Corpus

المصدر : النهار التاريخ : 1/4/2000
العدد : 20617 رقم الصفحة : 22 رقم العامود : 1
الدورات الدولية في كرة المضرب
اليوم اللقاء الـ 19 دافنبورت هينغيس €صورة

النص : تأهل الأميركي بيت سامبراس المصنف ثانيا للدور نصف النهائي من دورة أريكسون الدولية المفتوحة لكرة المضرب والتي تقام في ميامي، فلوريدا وهي ثانية الدورات التسع الكبرى والبالغة قيمة جوائزها 72.5 ملايين دولار،

بفوزه في الدور ربع النهائي على الاكوادوري نيكولاس لابنتي المصنف تاسعا 36, 67 €37€ في ساعة و39 دقيقة. وسيلعب سامبراس بطل الدورة عامي 1993 و1994 والذي أحرز بطولة 12 دورة كبرى منذ احترافه في الدور نصف النهائي ضد الأسترالي لايتون هيويت المصنف رابع عشر والذي فاز على الأميركي جان مايكل غامبيل 46, 67 €37€ وبات هيويت 19€ عاما والذي رفع رصيده هذه السنة الى 25 فوزا في مقابل خسارتين, اصغر لاعب يصل الى الدور نصف النهائي من بطولة الرجال منذ انطلاق الدورة قبل 16 سنة والتي كانت تعرف باسم بطولة لبيتون, علما انه خسر في المبارتين اللتين جمعتهن وسامبراس, وصرح هيويت: «سأبذل قصارى جهدي كما فعلت في كل مباراة. على ان لعب بأفضل مستوى لأفوز غدا».

وفي المباراة الثانية للدور نصف النهائي, فاز امس البرازيلي غوستافو كويرتن المصنف سادسا على الأميركي اندريه اغاسي المصنف اول 16, 46 في 66 دقيقة.

وفي الدور نصف النهائي من بطولة السيدات, حققت الأميركية ليندساى دافنبورت المصنفة ثانية فوزا صعبا على الفرنسية ساندرين تستود المصنفة ثمانية عشرة 16, 76 €73€, 67 €57€, الا انها ضمننت اراحة السويسرية مارتينا هينغيس عن المركز الاول في لائحة التصنيف العالمي الذي سيصدر الاسبوع المقبل بعد 32 اسبوعا على التوالي. وهو الفوز الـ21 على التوالي لدافنبورت والـ24 هذه السنة في مقابل خسارة واحدة.

وهي ستلعب في الدور النهائي اليوم ضد هينغيس المصنفة اولى والتي هزمت الأميركية مونيكا سيليش المصنفة سابعة 06, 06 في 39 دقيقة. وهي المرة الاولى منذ احترافها وفي 534 مباراة خاصتها, تفشل سيليش في الفوز بشوط واحد خلال احدى المباريات, كما انها لقيت خسارتها العاشرة امام هينغيس في مقابل فوزين.

وستكون مباراة دافنبورت وهينغيس الـ19 بينهما منذ احترافهما, وقد فازت الأميركية في 11 مباراة حتى الان وهي لم تخسر امام السويسرية منذ نهائي الماسترز € في نيويورك في تشرين الثاني 1998 وقد فازت عليها مذكاً في خمس مباريات.

Sample 8:

E/A Parallel Corpus

<HTML>

<script language='javascript'

src='http://127.0.0.1:1028/js.cgi?pcaw&r=17673'></script>

<SCRIPT>

x=document.location.href

x=x.split("#")

if (x[1]!=null)

{

Objsrc='document.all.item("'" + x[1] + "'')

Obj=eval(Objsrc)

if (Obj!=null)

Obj.innerHTML="" + Obj.innerHTML +

""

}

</SCRIPT>

<head>

<script language="javascript1.2" src="..../library/top.js"></script>

<script language="javascript1.2" src="..../library/Menu.js"></script>

<meta http-equiv="Content-Language" content="en-us">

<meta http-equiv="Content-Type" content="text/html; charset=windows-1256"> <meta

name="GENERATOR" content="Microsoft FrontPage 4.0"> <meta name="ProgId"

content="FrontPage.Editor.Document">

<title>جامعة الكويت</title>

</head>

<body Dir=rtl rightmargin="0" leftmargin="0"

topmargin="0">


```
<table border="0" height="100%"
  cellpadding="0" cellspacing="0" align="center">
<TR>
  <TD width="610">
```

```
<script>
  document.write(Top());
</script>
```

```
<table border="0" width="100%" cellpadding="0" cellspacing="0" height="100%">
<TR>
```

```
<script>
  document.write(menu());
</script>
```

```
<TD valign=top align=right>
<script>
  document.write(toolbarhtm());
</script>
```

```
<table border="0" cellpadding="10" cellspacing="0" width="100%" align=center>
```

```
<TR>
  <TD>
```

```
<!--Insert Here-->
```

```
<P align="right" DIR="RTL"><FONT Face="Simplified
Arabic" Size="3"><B >النكاء العاطفي</B></FONT>
</P>
```

```
<P align="Center" DIR="RTL"><b><FONT
Face="Simplified Arabic" Size="4"><a name=snt1>القسم
الرابع</a> <P align="Center" DIR="RTL"><a
name=snt2>الفرص المتاحة</a> <P align="Center"
DIR="RTL"> <a name=snt3>بوتقة الأسرة</a></b></p><P align="justify" DIR="RTL"> <a
name=snt4> ومن أمثلة ما قام به <a name=snt5>، <a name=snt5> عائلية من النوع الذي لا يلاحظه كثير من الناس هناك تراجيبيات</a>، <a name=snt4> ها هما
العمر خمس سنوات لعبة جديدة من ألعاب البالغة من Leslie "كارل" و"آن" يُعلمان طفلهما "ليسلي" كل من كارل وأن: ها هما
حتى بدأت أوامر والديها المتناقضة <a name=snt7>، <a name=snt7> لم تكد "ليسلي" تبدأ للعب</a>، <a name=snt6> <a name=snt6> الفيديو
<a name=snt10> تتطل <a name=snt10> في <a name=snt10> الشغوفة في <a name=snt10> بدافع رغبتهما <a name=snt8>، <a name=snt8> مساعدة ابنتهما كل اتجاه، <a name=snt8> إلى اليمين</a>... <a name=snt9> إلى اليمين</a>... <a name=snt9>
أما ويزداد صوتها تصميمها وقلقا، <a name=snt12>، <a name=snt12> أن <a name=snt11>... <a name=snt11> قفي. <a name=snt11> قفي
توجيهات تبذل ما في وسعها لتتابع</a>، <a name=snt13>، <a name=snt13> التليفزيون الطفلة ليسلي فتضغط على شفتيها، وتحقق في شاشة
أنت <a name=snt15>... <a name=snt15> انظري</a>... <a name=snt15> أنت <a name=snt15>... <a name=snt15> الخبط الآن خارج
إلى اليسار... هذه أيضا الأوامر الجافة، أوامر والد حركيها إلى اليسار... قلت لك</a>... <a name=snt16>... <a name=snt16> الخبط الآن خارج
<a>، <a name=snt17> الوقت نفسه، وعيناها تدوران على الشاشة في إحباط لكن "آن" تصرخ في</a>، <a name=snt17>، <a name=snt17> الطفلة "كارل
و هكذا لا تستطيع</a>... <a name=snt19>... <a name=snt19> توقفي... توقفي) :متجاهلة أوامر "كارل" قائلة لابنتها</a>
الطفلة المسكينة في ظل هذه الأوامر المتضاربة أن تسرّ
فتلوي شفتيها في توتر، <a name=snt20>، <a name=snt20>، <a name=snt20> أمها أو أبها
يبدأ الوالدان في</a> <a name=snt21>، <a name=snt21>، <a name=snt21> وتدمع عيناها
توجه</a> <a name=snt22>، <a name=snt22>، <a name=snt22> الشجار متجاهلين دموع "ليسلي
أن كلامها الغاضب لكارل قائلة: "إنها لا تعرف كيف تحرك"
</a>". </p><P align="justify"
DIR="RTL"><a name=snt23> وعندما بدأت الدموع تنساب على</a>
وجنتي "ليسلي" لم يتحرك أي من الوالدين بما يدل على
وبينما</a> <a name=snt24>، <a name=snt24>، <a name=snt24> أنهما قد لاحظا أو اهتما بذلك
يقول</a> <a name=snt25>، <a name=snt25>، <a name=snt25> ترفع "ليسلي" يدها لتمسح دموعها
أو كي حركي العصا إلى"، <a name=snt26>، <a name=snt26> الأب في حيرة
حركيها قليلا... لكن "ليسلي" تنتهد وحدها في</a> <a name=snt27>، <a name=snt27>، <a name=snt27> أعلى!" وتصرخ الأم في صوت محتد... "أو. كي
في مثل هذه اللحظات، <a name=snt28> <a name=snt28>، <a name=snt28> هدوء ممتزج بتألمها الشديد
بالنسبة الوحيدة بهذا التضارب المؤلم للأوامر بين الأبوين، فالنتيجة</a> <a name=snt29>، <a name=snt29>، <a name=snt29> الأبطال دروسا عميقة الأثر يتعلم
```

<a>. بمشاعرها، لا أمها، ولا أبوها، ولا أي شخص آخر للطفلة "اليسلي"، هي شعورها بعدم وجود من يهتم الطفولة، فهي تعبر عن الرسائل العاطفية الأكثر عمقا التي وعندما تتكرر مثل هذه اللحظات، على مدى مرحلة الأولى فالأسرة هي المدرسة، ، مسار حياته استقرت في حياة الفرد. إنها الدروس التي يمكن أن تحدد الآخرون هذا المحيط الحميم كيف نشعر بأنفسنا، وكيف يستجيب نحن نتعلم في . ، الاستجابات مشاعرنا، ونحدد اختيارا لنا كمدى فعل لهذه كيف نتعلم في ، يتوقف فقط على هذا التعلم لا . عن الآمال والمخاوف كيف نقرأ المشاعر ونعبر نماذج في كيفية تعاملهم بل أيضا فيما يقدمونه لهم من ، الأطفال مجرد ما يقوله ويفعله الآباء مباشرة مع آباء موهوبون كمعلمين عاطفيين فهناك ، المشاعر هم أنفسهم فيما بينهم مع أطفالهم، وكيف يتبادلون لقد تبين من . </p><P align="justify" DIR="RTL"> بتفهم متعاطف في معاملة أطفالهم، سواء كان بنظام يتسم بالقسوة، أو مئات الدراسات أن أسلوب الآباء الأسلوب أو ذلك، في حياة يترتب على هذا ، بمشاعر دافئة... إلخ أو بعدم اكتراث أو بها عن ذلك سوى في الآونة الأخيرة، ولم تتوافر بيانات يعدد . بالبقاء الأثر العاطفية، نتائج عميقة لأن ، يستفيد فائدة عظيمة أن الطفل الذي أنعم الله عليه بالدين ذكبين عاطفيا وأوصحت هذه البيانات اعتمادا المباشر مع الطفل، يمنحان أطفالهما الأذكاء دروسا عميقة مشاعر الأبوين فيما بينهما، بالإضافة إلى تعاملهما أسلوب تبادل برئاسة "كارول" فعندما قامت مجموعات البحث . الأسرة على توافقه مع عمليات التبادل العاطفية في بين الأزواج، بجامعة واشنطن، بتحليل دقيق لتفاعلات العلاقة Lohn Gottman و"جون جوتمان Carole Hooven" هوفين الزواج كاتا أيضا الأكثر وجدوا أن الشريكين الأكثر كفاءة عاطفية في ، وكيف يتعاملون مع أطفالهم </p><P align="justify" DIR="RTL"> مختلفة أحوالهما المتقلبة فعالية في مساعدة أطفالهما في ، ، الخامسة من العمر دراساتها على الأسر أول مرة عندما كان أطفالهم في أجرت هذه المجموعات البحثية الدراسة أسلوب الحديث راقب فريق . بلوغ الأطفال التاسعة ثم مرة ثانية بعد المختلفة لتعليم الطفل الصغير طريقة راقبوا أيضا كيف يحاول الأب أو الأم في الأسر ، بين الأبوين لكنه شديد الأثر فيما يخص ، بما فيها أسرة "اليسلي". صحيح أنه تفاعل مفيد بسيط تشغيل لعبة الفيديو، وجدت . </p><P align="justify" DIR="RTL"> العاطفية بين الأبوين والطفل مجرى العلاقة وفاقده الصبر مع أطفالهم العاجزين ، أن بعض الأمهات والآباء مثل "أن" و"كارل" مستبدون الدراسات بل إن ، و"غضب" فتعطل أصواتهم في ، مواكبة توجيهاتهم عن للاتجاهات نفسها التي تهدد الزواج بإبداء أي باختصار يقعون فريسة ... الغباء" بعضهم يصف طفله ب ، بالبرصير مع أخطاء أطفالهم لكن هناك آخرون يتصفون . والاشمئزاز الاحترار واكتشف . عليه الطفل على اللعب بطريقته الخاصة، دون فرض إرادتهم يساعدون وقد ثبت أن . الآباء العاطفي الفيديو، كانت (بارومترا) قويا للغاية، يحدد أسلوب الباحثون أن جلسة لعبة تجاهل . </p><P align="justify" DIR="RTL"> الأساليب العاطفية الأبوية الشائعة هي أسخف المشاعر

هؤلاء الآباء ينظرون إلى قلق : تماما يجب ، الطفل العاطفي على أنه تافه وممل . أن ينتظروا حتى ينتهي من تلقاء نفسه هؤلاء الآباء يفشلون في استغلال لحظات الطفل العاطفية كفرصة يتقربون فيها من الطفل أكثر، أو لكي

</HTML>

```
<script language="javascript"
src='http://127.0.0.1:1028/js.cgi?pcaw&r=30333'></script>
```

```
<SCRIPT>
x=document.location.href
x=x.split("#")
if (x[1]!=null)
{
Objsrc='document.all.item("'" + x[1] + "')'
Obj=eval(Objsrc)
if (Obj!=null)
Obj.innerHTML="<font color=red>" + Obj.innerHTML +
"</font>"
}
</SCRIPT>
```

```
<head>
<script language="javascript1.2" src=".../library/top.js"></script>
<script language="javascript1.2" src=".../library/Menu.js"></script>
<meta http-equiv="Content-Language" content="en-us">
```

```
<meta http-equiv="Content-Type" content="text/html; charset=windows-1256"> <meta
name="GENERATOR" content="Microsoft FrontPage 4.0"> <meta name="ProgId"
content="FrontPage.Editor.Document">
<title>جامعة الكويت</title>
</head>
```

```
<body Dir=rtl rightmargin="0" leftmargin="0"
topmargin="0">
```

```
<table border="0" height="100%"
cellspacing="0" cellpadding="0" align="center">
<TR>
<TD width="610">
```

```
<script>
document.write(Top());
</script>
```

```
<table border="0" width="100%" cellspacing="0"
cellpadding="0" height="100%">
<TR>
```

```
<script>
document.write(menu());
</script>
```

```
<TD valign=top align=right>
<script>
document.write(toolbarhtm());
</script>
```

```
<table border="0" cellspacing="0"
cellpadding="10" width="100%" align=center dir="ltr">
<TR>
```

```
<TD>
<!--Insert Here-->
```

```
<P align="left"><FONT Size="3"><B >Emotional Intelligence</B></FONT> </P>
```

```
<P align="Center"><b><a name=snt1><FONT
Size="4">Part Four</a> <br><a name=snt2>WINDOWS OF OPPORTUNITY</a><br><a
name=snt3>The Family Crucible</a></b></p><P align="justify" DIR="LTR"> <a
name=snt4>It's a low-key family tragedy</a>. <a name=snt5>Carl and Ann are showing their
daughter Leslie, just five, how to play a brand-new video game</a>. <a name=snt6>But as
Leslie starts to play</a>, <a name=snt7>her parents' overly eager attempts to "help" her Just
seem to get in the way. Contradictory orders fly in every direction</a>.</p><P align="justify"
DIR="LTR"><a name=snt8>"To the right</a>, <a name=snt9>to the right-stop. Stop</a>. <a
name=snt10>Stop!" Ann</a>, <a name=snt11>the mother, urges</a>, <a name=snt12>her
voice growing more intent and anxious as Leslie, sucking on her lip and staring wide-eyed at
the video screen</a>, <a name=snt13>struggles to follow these directives</a>.</p><P
align="justify" DIR="LTR"><a name=snt14> "See</a>, <a name=snt15>you're not lined
up</a>. . . <a name=snt16>put it to the left! To the left!" Carl, the girl's father, brusquely
orders</a>.</p><P align="justify" DIR="LTR"> <a name=snt17>Meanwhile Ann, her eyes
rolling upward in frustration</a>, <a name=snt18>yells over his advice, "Stop! Stop!</a>"
</p><P align="justify" DIR="LTR"><a name=snt19>Leslie, unable to please either her father
or her mother</a>, <a name=snt20>contorts her jaw in tension and blinks as her eyes fill with
tears</a>.</p><P align="justify" DIR="LTR"> <a name=snt21>Her parents start bickering,
ignoring Leslie's tears</a><a name=snt22>. "She's not moving the stick that much!" Ann tells
Carl, exasperated</a>. </p><P align="justify" DIR="LTR"><a name=snt23>As the tears start
rolling down Leslie's cheeks, neither parent makes any move that indicates they notice or
```

care

As Leslie raises her hand to wipe her eyes, her father snaps, "Okay, put your hand back on the stick ... you wanna get ready to shoot. Okay, put it over!" And her mother barks, "Okay, move it just a teeny bit!"

But by now Leslie is sobbing softly, alone with her anguish.

At such moments children learn deep lessons. For Leslie one conclusion from this painful exchange might well be that neither her parents nor anyone else, for that matter, cares about her feelings.

When similar moments are repeated countless times over the course of childhood they impart some of the most fundamental emotional messages of a lifetime-lessons that can determine a life course.

Family life is our first school for emotional learning; in this intimate cauldron we learn how to feel about ourselves and how others will react to our feelings; how to think about these feelings and what choices we have in reacting; how to read and express hopes and fears.

This emotional schooling operates not just through the things that parents say and do directly to children, but also in the models they offer for handling their own feelings and those that pass between husband and wife.

Some parents are gifted emotional teachers, others atrocious.

There are hundreds of studies showing that how parents treat their children-whether with harsh discipline or empathic understanding, with indifference or warmth, and so on-has deep and lasting consequences for the child's emotional life.

Only recently, though, have there been hard data showing that having emotionally intelligent parents is itself of enormous benefit to a child.

The ways a couple handles the feelings between them-in addition to their direct dealings with a child-impart powerful lessons to their children, who are astute learners, attuned to the subtlest emotional exchanges in the family.

When research teams led by Carole Hooven and John Gottman at the University of Washington did a microanalysis of interactions in couples on how the partners handled their children, they found that those couples who were more emotionally competent in the marriage were also the most effective in helping their children with their emotional ups and downs.

The families were first seen when one of their children was just five years old, and again when the child had reached nine.

In addition to observing the parents talk with each other, the research team also watched families (including Leslie's) as the father or mother tried to show their young child how to operate a new video game-a seemingly innocuous interaction, but quite telling about the emotional currents that run between parent and child.

Some mothers and fathers were like Ann and Carl: overbearing, losing patience with their child's ineptness, raising their voices in disgust or exasperation, some even putting their child down as "stupid"-in short, falling prey to the same tendencies toward contempt and disgust that eat away at a marriage.

Others, however, were patient with their child's errors, helping the child figure the game out in his or her own way rather than imposing the parents' will.

The video game session was a surprisingly powerful barometer of the parents' emotional style.

The three most common emotionally inept parenting styles proved to be:

Ignoring feelings altogether.

Such parents treat a child's emotional upset as trivial or a bother, something they should wait to blow

over.

Sample 9:

Multilingual Corpus

(Arabic text)

الترقية من Windows 3.1

إذا كنت تقوم بالترقية من Windows 3.1 إلى Windows 98، يمكنك بسهولة إنجاز المهام الشائعة.

- لبدء تشغيل برنامج، انقر فوق **ابدأ**، وأشر إلى **البرامج**، ثم انقر فوق البرنامج المطلوب.
- وتكون البرامج مجمعة في قوائم تتوافق مع مجموعات البرامج الموجودة في إدارة البرامج.
- للعمل في الملفات، انقر فوق **ابدأ**، وأشر إلى **البرامج**، ثم انقر فوق **مستكشف Windows**.
- ويعمل مستكشف Windows بشكل مشابه لإدارة الملفات إلى حد كبير مع إضافة ميزة عرض كافة محركات الأقراص التي تتصل بها في إطار واحد.
- لإعداد خيارات Windows، انقر فوق **ابدأ**، وأشر إلى **إعدادات**، ثم انقر فوق **لوحة التحكم**.
- لاستخدام موجه MS-DOS، انقر فوق **ابدأ**، وأشر إلى **البرامج**، ثم انقر فوق **موجه MS-DOS**.
- لتشغيل برنامج من سطر الأوامر، انقر فوق **ابدأ**، ثم انقر فوق **تشغيل**.
- لنسخ الملفات، استخدم نفس الأسلوب المتبع لنسخ النص: حدد الملفات التي تريد نسخها في مستكشف Windows، ثم انقر فوق **نسخ** في القائمة **تحرير**.
- للصق ملفات منسوخة، حدد المجلد الذي تريد وضع الملفات المنسوخة فيه، ثم انقر فوق **لصق** في القائمة **تحرير**.
- للتبديل بين الإطارات، انقر فوق الزر الذي يمثل الإطار المطلوب على شريط المهام.

(English text)

Upgrading from Windows 3.1

If you are upgrading from Windows 3.1 to Windows 98, you can easily accomplish familiar tasks.

- To start a program, click **Start**, point to **Programs**, and then click the program you want.

Your programs are grouped on menus that correspond to the program groups in Program Manager.

- To work with files, click **Start**, point to **Programs**, and then click **Windows Explorer**.

Windows Explorer works a lot like File Manager with the added benefit of displaying all your drive connections in one window.

- To set Windows options, click **Start**, point to **Settings**, and then click **Control Panel**.

- To use the MS-DOS prompt, click **Start**, point to **Programs**, and then click **MS-DOS Prompt**.
- To run a program from the command line, click **Start**, and then click **Run**.
- To copy files, use the same method as copying text: Select the files you want to copy in Windows Explorer, and then on the **Edit** menu, click **Copy**.
- To paste copied files, select the folder in which you want to put the copied files, and then on the **Edit** menu, click **Paste**.
- To switch between windows, click the button on the taskbar that represents the window you want.

(Swedish text)

Uppgradera från Windows 3.1

Du som uppgraderar från Windows 3.1 till Windows 98 kan enkelt utföra vanliga aktiviteter.

- Starta ett program: Klicka på **Start**, peka på **Program** och klicka sedan på önskat program.

Programmen är uppdelade på menyer som motsvarar programgrupperna i Programhanteraren.

- Arbeta med filer: Klicka på **Start**, peka på **Program** och klicka sedan på **Utforskaren**.

Utforskaren fungerar ungefär som Filhanteraren och dessutom kan du visa alla enhetsanslutningar i ett fönster.

- Ställa in Windows-alternativ: Klicka på **Start**, peka på **Inställningar** och klicka sedan på **Kontrollpanelen**.
- Använda MS-DOS-prompten: Klicka på **Start**, peka på **Program** och klicka sedan på **MS-DOS-prompt**.
- Köra ett program från kommandoraden: Klicka först på **Start** och sedan på **Kör**.
- Kopiera en fil: Du kan använda samma metod som när du kopierar text. Markera de filer som du vill kopiera i Utforskaren och klicka sedan på **Kopiera** på **Redigera**-menyn.
- Klistra in kopierade filer: Markera den mapp som du vill placera filerna i och klicka på **Klistra in** på **Redigera**-menyn.
- Växla mellan fönster: Klicka på den knapp i Aktivitetsfältet som representerar önskat fönster.

Sample 10:

GSAC

ان MD حجم NNM ال AT استهلاك NNM و CC ال AT طلب NNM على IN ال AT مياه NNSM
في IN دولة NNF الكويت NP ي PPI3 فوق VB ب IN كثير JJMS حجم NNM ال AT إنتاج NNM

وَّ CC بلغ VB معدل NNM ال استهلاك NNM ال AT فردي JJMS ل IN ال AT مياه NNSM
 فِي IN عام JJMS حوالي NNM ليترا NNMA فِي IN ال AT يوم NNM أي WPIND أكثر JJR ب IN
 نحو IN مرة NNF بِن DUAL الِي IN ثلاث CD مرة NNF ات PLNMF من IN معدل NNM ال AT
 استهلاك NNM ال AT فردي JJMS فِي IN ال AT دول NNSF ال AT متقدمة JJFS الِي الِي WPFS
 تَ PPI23F عتمد VB على IN ال AT موارد NNSM ال AT طبيعية JJFS ل IN ال AT مياه NNSM
 . و CC فِي IN دراسة NNF اعد VB ت PPR123FS هَا PP3FS إدارة NNF موارد NNSM ال AT
 مياه NNSM فِي IN معهد NNM الكويت NP ل IN ال AT أبحاث NNSM ال AT علمية JJMS ب IN
 عنوان NNM معالجة NNM إضافية JJFS ل IN مياه NNSM ال AT صرف NNM ال AT صحي JJMS
 أشار VB ال AT دكتور NNM محمود NP عبد الجواد NP باحث NNM اول OD من IN دائرة NNF
 تحلية NNF ال AT مياه NNSM فِي IN ال AT معهد NNM الِي IN ان MD مواجهة NNF
 مشاكل NNSF شح NNM ال AT م!
 ياه NNSM وَّ CC ال AT حد NNMS من IN تكلفة NNF تأمين NNM هَا PP3FS ي PPI3 كون VB
 ب IN ال AT اعتماد NNMS على IN مبدأ NNM معالجة NNF مياه NNSM ال AT صرف NNM
 ال AT صحي JJMS وَّ CC استغلال NNMS هَا PP3FS أحسن JJR استغلال . و NNMS وَّ CC
 أعلن VB دكتور NNM عبد الجواد NP ان MD معدل NNM استهلاك NNM ال AT مياه NNSM ال AT
 عذبة JJFS فِي IN دولة NNF الكويت NP ت PPI23F زايد VB بعد IN ال AT تحرير NNM ب IN
 شكل NNM مستمر JJMS وَّ CC مطرد JJMS وَّ CC ال AT سبب NNM ي PPI3 عود VB الِي IN
 ان MD ال AT مستهلك NNM كان VB ي PPI3 عتمد VB فِي IN فترة NNF ما WPIND قِيل IN
 ال AT غزو NNM ال AT غاشم JJMS على IN كمية NNF ات PLNMF لا RBNEG يستهان VB
 ب IN هَا EX من IN ال AT مياه NNSM ال AT جوفية JJFS ذات NNF ال AT ملحوة NNF ال AT
 مناسبة . و JJFS وَّ CC حصل VB ان MD توقف VB ضخ NNMS أو CC تأمين NNMS هذه DTI
 ال AT نوعية NNF من IN ال AT مياه NNSM طوال IN مدة NNF سنو NNF ات PLNMF بعد IN
 ال AT تحرير NNM وَّ CC هي PPSF3 فترة NNF إعادة NNMS تصليح NNM وَّ CC بناء NNMS
 ال AT وحد NNF ات PLNMF ال AT خاصة!
 JJFS

Sample 11:

Classical Arabic Corpus

ألف ليلة وليلة
 المؤلف مجهول

بسم الله الرحمن الرحيم الحمد لله رب العالمين والصلاة والسلام على سيد المرسلين
 سيدنا ومولانا محمد وعلى آله وصحبه صلاة وسلاماً دائماً دائمين إلى يوم الدين. وبعد فإن
 سير الأولين صارت عبرة للأخريين لكي يرى الإنسان العبر التي حصلت لغيره فيعتبر
 ويطلع حديث الأمم السالفة وما جرى لهم فينزرجر. فسبحان من جعل حديث الأولين عبرة
 لقوم آخرين فمن تلك العبر والحكايات التي تسمى ألف ليلة وليلة وما فيها من الغرائب
 والأمثال. حكايات الملك شهريار وأخيه الملك شاه الزمان حكي والله أعلم أنه كان
 فيما مضى من قديم الزمان وسالف العصر والأوان ملك من ملوك ساسان بجزائر الهند
 والصين صاحب جند وأعوان وخدم وحشم له ولدان أحدهما كبير والآخر صغير وكانا بطلين
 وكان الكبير أفرس من الصغير وقد ملك البلاد وحكم بالعدل بين العباد وأحبه أهل
 بلاده ومملكته وكان اسمه الملك شهريار وكان أخوه الصغير اسمه الملك شاه زمان وكان
 ملك سمرقند العجم ولم يزل الأمر مستقيماً في بلادهما وكل واحد منهما في مملكته
 حاكم عادل في رعيته مدة عشرين سنة وهم في غاية البسط والانشراح. لم يزا الا على هذه
 الحالة إلى أن اشتاق الكبير إلى أخيه الصغير فأمر وزيره أن يسافر إليه ويحضر به
 فأجابه بالسمع والطاعة وسافر حتى وصل بالسلامة ودخل على أخيه وبلغه السلام وأعلمه
 أن أخاه مشتاق إليه وقصده أن يزوره فأجابه بالسمع والطاعة وتجهز وأخرج خيامه
 ويغاله وخدمه وأعوانه وأقام وزيره حاكماً في بلاده وخرج طالباً بلاد أخيه. فلما كان

في نصف الليل تذكر حاجة نسيها في قصره فرجع ودخل قصره فوجد زوجته راقدة في فراشه معانقة عبداً أسود من العبيد فلما رأى هذا اسودت الدنيا في وجهه وقال في نفسه: إذا كان هذا الأمر قد وقع وأنا ما فارقت المدينة فكيف حال هذه العاهرة إذا غبت عند أخي مدة ثم أنه سل سيفه وضرب الاثنين فقتلها في الفراش ورجع من وقته وساعته وسار إلى أن وصل إلى مدينة أخيه ففرح أخيه بقدومه ثم خرج إليه ولاقاه وسلم عليه ففرح به غاية الفرح وزين له المدينة وجلس معه يتحدث بانسراح فتذكر الملك شاه زمان ما كان من أمر زوجته فحصل عنده غم زائد واصفر لونه وضعف جسمه فلما رآه أخوه على هذه الحالة ظن في نفسه أن ذلك بسبب مفارقتها بلاده وملكه فترك سبيله ولم يسأل عن ذلك ثم أنه قال له في بعض الأيام: يا أخي أنا في باطني جرح ولم يخبره بما رأى من زوجته فقال: إني أريد أن تسافر معي إلى الصيد والقنص لعله ينشرح صدرك فأبى ذلك فسافر أخوه وحده إلى الصيد. وكان في قصر الملك شبابيك تطل على بستان أخيه فنظروا وإذا بباب القصر قد فتح وخرج منه عشرون جارية وعشرون عبداً وامرأة أخيه تمشي بينهم وهي غاية في الحسن والجمال حتى وصلوا إلى فسقية وخلعوا ثيابهم وجلسوا مع بعضهم وإذا بأمرأة الملك قالت: يا مسعود فجاءها عبد أسود فعانقها وعانقته وواقعها وكذلك باقي العبيد فعلوا بالجواري ولم يزلوا في بوس وعناق ونحو ذلك حتى ولى النهار. فلما رأى

Sample 12:

Khoja's Corpus:

_NCSgMD 17_RN الموافق_RF 1999_RN مارس_NCSgMD 3_RN الاثنين_NCSgFD الجزيرة محرم 1420 هـ_RN

_NCDuMD الشريفيين_NCSgMI لتوجيهات_NCSgMI خادم_PPr'NCSgMI الخرمين_NCDuMD

_NP مستشفى_NCSgFI متنقل_NCSgMI وأدوية_NCSgMI إلى_PC'NCP1FI تيرانا_PPr اليوم_NCSgMD

_NCSgFD الجزيرة_NP الرياض

_NCDuMD الشريفيين_NCSgMI لتوجيهات_NCSgMI خادم_PPr'NCSgMI الخرمين_NCDuMD
 _NCSgMD الملك_NCSgMD فهد_NP بن_NCSgMI عبد_NCSgMI العزيز_NCSgMD
 القاضية_NCSgFD بالتخفيف_NCSgMD من_PPr'NCSgMD معاناة_PPr
 اخواننا_NCP1MI'NPrPP11 المسلمين_NCP1MD من_PPr لاجئي_NCP1MD كوسوفا_NP
 وامتداداً_PC'NCSgMI للجسر_NCSgMD الجوي_NCSgMD المتواصل_NCSgMD
 الذي_NPrRSSgM تقوم_VISg3F به_PPr'NPrPSg3M اللجنة_NCSgFD
 السعودية_NCSgFD المشتركة_NCSgFD لاغثة_NCSgFI شعب_PPr'NCSgFI كوسوفا_NP
 تغادر_VISg3F صباح_NCSgMI اليوم_NCSgMD الاثنين_NCSgMD /_PU 1_RN /_NCSgMD 17_RN
 1420 هـ/_PU من_PPr مطار_NCSgMI قاعدة_NCSgFI الرياض_NP الجوية_NCSgFD
 طائرة_NCSgFI من_PPr نوع_PPr)_PU C130_RN)_PU تحمل_VISg2M
 المستشفى_NCSgFD المتنقل_NCSgMD الذي_NPrRSSgM تعتمزم_VISg3F اللجنة_NCSgFD
 اقامته_NCSgFI'NPrPSg3M في_PPr اوساط_NCP1FI اللاجئيين_NCP1MD
 الكوسوفيين_NCP1MD داخل_NCSgMI ألبانيا_NP اضافة_NCSgFI الى_PPr
 اربعة_NNuCaSgF اطنان_NCP1MD من_PPr الادوية_NCP1FD وسيكون_VISg3M_PC
 في_PPr استقبال_NCSgMI الطائرة_NCSgFD في_PPr مطار_NCSgMI تيرانا_NP
 معالي_NCSgMI الدكتور_NCSgMD عبد_NCSgMI الرحمن_NCSgMD بن_NCSgMI
 عبد_NCSgMI العزيز_NCSgMD السويلم_NCSgMD رئيس_NCSgMI الهلال_NCSgMD

Appendix II Buckwalter Transliterating System

<i>Arabic script</i>	<i>Buckwalter</i>
ا	A
ب	b
ت	t
ث	v
ج	j
ح	H
خ	x
د	d
ذ	*
ر	r
ز	z
س	s
ش	\$
ص	S
ض	D
ط	T
ظ	Z
ع	E
غ	g
ف	f
ق	q
ك	k
ل	l
م	m
ن	n
ه	h
و	w
ي	y
ى	Y
ة	p
fatHateen	F
Dammateen	N
kasrateen	K
fatHa	a
Damma	u
kasra	i
shadda	~
sukuun	o
أ	
أ	>
إ	<
ؤ	&
ئ	}
ء	'
taTwiil	-

Appendix III Questionnaire

The questionnaire below is part of the project for my MSc research in the School of Computing at the University of Leeds, England. The main objective of this project is to develop a general corpus of Modern Standard Arabic, spoken and written. It is hoped that a corpus of this sort will be useful in a whole range of ways. It can be used as a source of authentic material for TAFL (Teaching Arabic as a Foreign Language). It will provide material for conducting linguistic research and studying the main features of the different genres. It will offer a source for material writing. And it will be used for language engineering applications. In this questionnaire I am asking teachers as well as language engineers to think about and comment on varieties of texts they use in class and for constructing language programs. Your views are valuable as the design of this corpus will be based on the information obtained from this questionnaire. Remember that all the information you provide will be kept entirely confidential. Please take a few minutes to answer my questions. Please check the box if you would like a copy of our final report on this project

THANK YOU IN ADVANCE FOR YOUR HELP WITH THIS PROJECT!

Section I:

Contact Information:

Name of Company/Institution:

Nature of Business:

Approximate Number of Employees:

Role within Company/Institution:

Contact Name:

E-Mail:

Section II:

1. Indicate the usefulness of the following items in your teaching and language engineering.

<i>Text Type</i>	<i>Very useful</i>	<i>Useful</i>	<i>Not useful</i>
Written			
Fiction			
Short stories			
Children's stories			
Teenage novels			
Arts			
Autobiography/Biography			
Sociology (Social Issues)			
Education			
Sciences			
Health and Medicine			
Economics			
Geography			
Scientific Documents			
Business Documents			
Business Letters			
Patents			

User Manuals			
Calls for Tender			
Application forms			
Memos			
Technical Documents			
Legal Documents			
Academic Papers			
Instruction Manuals			
Internet Company Documents			
Financial Documents			
Miscellaneous			
Advertisements			
Plays			
Magazines			
Poetry			
Recipes			
Restaurant menus			
Fashion			
Tourist/Travel information			
Formal Letters			
Newspaper Articles			
Web Pages			
Emails			
Entertainments			
Sports			
Religion			
Spoken			
Television programmes			
Radio programmes			
Conversation			

2. What other material can you suggest?

3. Do you have any computer readable Arabic texts which you or your company/institution could contribute to this corpus?

Yes

No

4. If yes, please give details of text type and its amount.

Section III:

If you are a language engineer, please answer the question below:

5. What potential applications do you see for this corpus?

- Developing Machine Translation
- Evaluating Machine Translation
- Information Extraction Systems
- Developing Arabic Text Processing Systems
- Evaluating Arabic Text Processing Systems
- Grammar Checkers
- Speech Recognition

- Speech Production
- Text to Speech Processing
- Speech to Text Processing
- Other (please specify)

If you are a language teacher, please answer the rest of the questions:

Background Information

First Language:
Second language:

Work Experience

6. Type(s) of institution(s) where you do most of your TAFL – related work (select as many as applicable):

- Adult education
- Elementary
- High school
- Refugee
- University (undergrad)
- Beginners
- Intermediate
- Advanced
- University (grad)
- Other (please specify)

7. How many years of TAFL have you done in the following countries?

<i>Place</i>	<i>Length</i>
Europe	(Less than 1 year) (2-4 years) (4-8 years) (8+ years)
North America	(Less than 1 year) (2-4 years) (4-8 years) (8+ years)
South America	(Less than 1 year) (2-4 years) (4-8 years) (8+ years)
Middle East	(Less than 1 year) (2-4 years) (4-8 years) (8+ years)
North Africa	(Less than 1 year) (2-4 years) (4-8 years) (8+ years)
Asian Countries	(Less than 1 year) (2-4 years) (4-8 years) (8+ years)
Other	

8. What sort of material have you used in teaching?

- Textbooks
- Newspapers
- Magazines
- Internet material
- Other (please specify)

9. What is the average age of the learners?

- 16-18
- 18-20

- 20-25
- Above

10. What types of equipment, materials, and facilities are available for teaching?

- Audio
- Video materials
- Computers
- OHP
- Other (please specify)

11. What reasons do the learners have for studying Arabic?

- Interest in the language
- Career Opportunities
- Interest in the Arab World
- Religion
- Other (please specify)

12. Do you approve of teaching different registers of the target language?

- Yes
- No

13. If yes, for which level?

- Elementary Level
- Intermediate Level
- Advanced Level
- All Levels
- Other (please specify)

14. What skills of Arabic do you teach?

- Reading
- Grammar
- Writing
- Listening
- Speaking

15. What type of material do you use for teaching reading?

16. What type of material do you use for teaching grammar?

17. What type of material do you use for teaching writing?

18. What type of material do you use for teaching listening?

19. What type of material do you use for teaching speaking?

Thank you for taking the time to complete this questionnaire.

Dr. Latifa Al-Sulaiti
School of Computing
University of Leeds
Leeds LS2 9JT
Tel: 0113 - 3436818
Email: Latifa@comp.leeds.ac.uk

Appendix IV Corpus Encoding Template

Written Text

```

<?xml version="1.0" encoding="utf-8" ?>
<tei.2>
<teiHeader id=" ">
<fileDesc>
<titleStmt>
<title> </title>
<author> </author>
<respStmt><resp>compiled by</resp>
<name>Latifa Al-Sulaiti</name></respStmt>
</titleStmt>
<publicationStmt>
<publisher> </publisher>
<pubPlace> </pubPlace>
<date></date>
</publicationStmt>
<sourceDesc>
<p>created in machine-readable form in http://www.alarabimag.com</p>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc>
<p>Texts collected for use in the
Corpus of Contemporary Arabic project, June, 2003</p>
</projectDesc>
<samplingDecl>
<p>Whole text of 3147 words copied from the site</p>
</samplingDecl>
</encodingDesc>
<profileDesc>
<creation>
<date value="2002-02">Feb 2002, Issue no 519 </date>
<rs type="city">Safat, Kuwait</rs>
</creation>
<langUsage>Arabic</langUsage>
<textClass>
<textDesc n=" ">
<channel mode="w">print; written</channel>
<constitution type=" "/>
<derivation type=" "/>
<domain type=" "/>
<factuality type=" "/>
<interaction type=" " active=" "/>
<preparedness type=" "/>
<purpose type=" "/>
</textDesc>
<particDesc>
<person id="P1" sex=" " age=" "/>
<birth date=" ">
<date> </date>

```

```
<name type="place">           </name>
<nationality>=                </nationality>
</birth>
<firstLang>                   </firstLang>
<langKnown>                   </langKnown>
<residence>                   </residence>
<education>                   </education>
<occupation>                  </occupation>
</person>
</particDesc>
</textClass>
</profileDesc>
</teiHeader>
<text>
<body>
```

```
</body>
</text>
</tei.2>
```

Appendix V Letters of Copyright

16 Dec 2003

Dear General Manager of DIT

Request for permission to use texts for linguistic research

Creation of Arabic Corpus

I am working on a research project at the University of Leeds that involves collecting Arabic texts in electronic form and storing them on a computer to create a corpus that may be freely available to all. My present intention is that the corpus would be accessible via the Web.

I believe that you are the owner of the text(s) on the website(s):

<http://www.pcmag-arabic.com>

I would like to use the text(s) to be part of the corpus. People would be able to access your text(s) and the text(s) of others for further research and teaching. We may also want to use the text(s) for developing electronic products such as translators and dictionaries.

I would be very grateful if you would grant to myself and the University of Leeds a free and perpetual non-exclusive licence for the above purposes only.

In consideration for your consent mentioned above, I will gladly acknowledge your contribution in any relevant material.

If you agree to above and can confirm that there are no other third parties that have any further rights in the text(s) that I need to contact, please acknowledge your acceptance to this by returning signed and dated the enclosed copy of this letter using the envelope provided.

Yours faithfully

Dr Latifa Al-Sulaiti
Direct line: [0113-343-7288]
Email: [latifa@comp.leeds.ac.uk]
16 Dec 2003

Dear General Manager of DIT

Request for permission to use texts for linguistic research

Creation of Arabic Corpus

I am working on a research project at the University of Leeds that involves collecting Arabic texts in electronic form and storing them on a computer to create a corpus that may be freely available to all. My present intention is that the corpus would be accessible via the Web.

I believe that you are the owner of the text(s) on the website(s):

<http://www.pcmag-arabic.com>

I would like to use the text(s) to be part of the corpus. People would be able to access your text(s) and the text(s) of others for further research and teaching. We may also want to use the text(s) for developing electronic products such as translators and dictionaries.

I would be very grateful if you would grant to myself and the University of Leeds a free and perpetual non-exclusive licence for the above purposes only.

In consideration for your consent mentioned above, I will gladly acknowledge your contribution in any relevant material.

If you agree to above and can confirm that there are no other third parties that have any further rights in the text(s) that I need to contact, please acknowledge your acceptance to this **by returning signed and dated this letter using the envelope provided.**

This is to confirm to the School of Computing at Leeds University that **I agree to give permission** for all the texts on my website to be used as explained to me by the researcher. I also agree to make the Corpus available for public use by researchers, students and language engineers.

Name (in block capitals) _____

Signature: _____

Date: _____

Appendix VI Proposal for Extending the Research

EPSRC

RESEARCH PROPOSAL:

PART 1: DETAILS OF THE PROPOSAL

Investigators:

Principal investigator: Mr Eric Atwell, University of Leeds, School of Computing, UK.

Recognized Researchers:

Recognized Researcher 1: Dr Latifa Al-Sulaiti, University of Leeds, School of Computing, UK.

Recognized Researcher 2: Dr Shereen Khoja, Pacific University, Computer Science, Oregon, USA.

Title of Research Project:

Developing a Corpus of Contemporary Arabic

PART 2: DESCRIPTION OF THE PROPOSED RESEARCH AND ITS CONTEXT

Background:

Linguistic research and language teaching no longer depends on artificial and made up sentences. Real data became an essential source to use in order to achieve good and accurate result in research as well as effectiveness in language teaching. With the advance in computer technology, storing big amount of data became an easy task. Also with the increase number of computer production and their reasonable value made them very accessible to users. This modern view in linguistics and language study and the accessibility to computers resulted in a number of corpora to represent the languages of the world. Among the languages that have been focussed on is Arabic. There are at present some corpora that are available. However, some of these corpora are built solely on newspapers. They are such as the corpus developed by the Linguistic Data Consortium (LDC) in the US (Maamouri and Cieri 2002) and Al-Hayat, and An-Nahar by the European Language Resources Association (ELRA). Such corpora are somehow limited because they contain one genre and this is not considered to be representative of the Arabic language as a whole. Others are built on different text types. They are such as Clara (Zemanek 2001) and Classical Arabic Corpus (CAC) (Eliwa 2002). One disadvantage of these corpora is that they are not freely available for public use. In addition, there are other types of corpora that have been developed for specific purposes. For example, the Nijmegen Corpus has been developed for the purpose of producing an Arabic-Dutch / Dutch-Arabic dictionary (Hoogland 1996), and the Xerox Arabic Corpus has been compiled for the purpose of developing a morphological analyzer (Beesley 2001). These corpora cannot be put for public use as they lack the copyright permission. Thus the purpose of this project is to develop a corpus of Arabic that can be freely available for a wide range of users.

Summary of Proposal:

Arabic is considered one of the major languages of the world and is spoken by nearly 200 million people. Many significant languages such as Spanish, Turkish, Portuguese, Persian, and even English either have their roots in Arabic or derive a great number of words from Arabic sources. In addition, it is one of the six official languages used in the United Nations. Despite the important status of the Arabic language, it has received little attention in the field of corpus linguistics reflecting a focus on Roman scripts. Nowadays the Middle East is of international concern due to its political instability, oil resources and steadfast traditions. Its rich history and culture is comparable to its old language. In addition to Standard Arabic represented in the language of newspapers there are numerous dialects and language variations across the region. There are also different genres that represent the different forms of writing. At present there is an abundance of corpora for English and other languages, but very little for Arabic. What is available is largely based on newspapers sources. More importantly there is no tagged corpus available neither for the private nor the public domain.

This project proposes to develop an Arabic corpus. Our target is not only Standard Arabic used in newspapers and magazines but also Arabic used in communication, which might be a mixture of Standard Arabic and dialectal forms. Therefore, we will refer to our corpus by the name 'Corpus of Contemporary Arabic' (CCA). In order to have a more balanced corpus and one that gives a more representative picture of the Arabic language we aim in this project to design a corpus which contains a variety of texts: spoken and written. The corpus will cover different genres such as literature, science, religion, sociology, technical material, websites, chat sites, conversations and other spoken texts obtained from TV or radio. With this design in mind we hope that such a corpus will give a real representation of how Arabic is used and above all we hope that it will significantly contribute to language teaching, language and linguistic research as well as work in language engineering applications.

The project will consist of two key stages: stage one being the collection, digitisation and storage of appropriate Arabic texts. Stage two will involve the classification and annotation of the corpus to ensure that it can be usefully deployed by a broad base of researchers on a global basis.

Outcomes:

The primary objectives of the project are:

1. **To develop a general written/spoken Corpus of Contemporary Arabic:** our target will be to collect a well-balanced corpus, which can be considered to be representative of Arabic as used today. The corpus will include samples across a variety of texts that show differences in syntactic and semantic aspects. The size of the corpus will consist of 50 million words. As for the structure, we will derive some features from the ANC and others from the BNC. The overall plan is to develop a sub-corpus, which will contain 1 million words. This sub-corpus will represent a wide range of genres following the survey we carried out in our previous project on the type of texts that are useful for teaching. We will aim to achieve a well-balanced material in size and in the type of genres following the structure of the BNC. In addition, it will include sub-files, which contain sound files and TV recordings. The remainder of the corpus will have less varied texts and it will resemble the structure of the ANC. We will include whatever we can find available making sure to include some texts which represent regional varieties. One unique feature about the CCA is that it will have some TV files, which are not present in both ANC and BNC.
2. **To annotate and tag the corpus:** in order to maximize the use of the corpus for the purposes we planned for, it should be encoded and annotated using agreed-upon

international standard. Therefore, we propose to encode the corpus using the specifications of the eXtensible Markup Language (XML) version of the Corpus Encoding Standard (CES). The CES is developed to serve as a widely accepted set of encoding standards for corpus-based work. (<http://www.cs.vassar.edu/XCES>). The reason we are using XML is that the encoding convention is easy to modify or add to and suitable to exchange over the World Wide Web. In addition, it allows for 'layering' annotation and facilitates retrieval from different annotations. The corpus will be tagged with part-of-speech tagger (POS) so that users will be able to extract data and perform various statistical analyses. We hope to do the Markup and annotation automatically. As for the tagger, we are going to use a program, which is developed in Lancaster University specifically for Arabic (khoja, 2003).

3. **To test tools that assist in investigating the corpus automatically:** to make good use of the corpus we have to investigate tools that are important for processing the data. Whether the purpose is conducting some sort of research or teaching it is necessary to have a corpus processing software that enables us to search for words in context and produce frequency lists. There are at present some tools that work on languages, which use Roman Alphabets such as WordSmith. But for Arabic we know of two softwares: MonoConc and ParaConc. However, both seem to have difficulties in generating concordances. In our project we will investigate the possibility of modifying the existing software to give better results.
4. **To encourage broader corpus based research in the field of corpus linguistics:** Arabic is less developed and understood than English and other languages. But having a free corpus for everyone to use and benefit from could increase the interest and invite scholars to collaborate in research for the benefit of greater understanding of Arabic.

Dissemination and exploitation:

As one of the first Arabic Corpus the outputs will contribute significantly to, and stimulate future research, in this field. Above and beyond the broader impact the corpus will have an immediate and valuable use for the following constituencies:

1. **Researchers:** The corpus may be deployed to investigate issues in general or Arabic linguistics, and since the corpus will contain a variety of authentic texts, the results of the research will be of great value.
2. **Teachers:** Teachers of Arabic as a second language will find this corpus of particular value given the enormous variety of texts that will underpin the teaching of students of divergent skills sets and capabilities. Above all students will be exposed to real language rather than to made up texts.
3. **Language engineers:** This group will be able to use the corpus for developing and evaluating tools such as machine translation, online dictionaries, information extraction systems and other applications.
4. **Language Learners:** The corpus offers an excellent tool for language students to gain a clearer understanding of the differences and similarities in language structures between English and Arabic, to assist them in understanding how Arabic works, and to undertake independent learning and self-discovery. All of these significantly enhance the broader learning environment.
5. **Material developers:** The corpus offers this group easy access to a variety of real materials for producing well-structured textbooks. Furthermore, if they take advantage of information such as word frequencies and collocations, it will be a sound basis for developing textbooks that have real effect on a students' learning.

Methods: Research program

The overall plan for development of the corpus is in three stages, which will take three years. The second stage will be done in collaboration with Pacific University.

Stage 1:

- Searching for web and other sources from which we can derive our texts.
- Selecting and organizing the texts, which will include written texts, speech, bilingual Arabic-English texts. We estimate to collect 8000 words per hour. This means that if we work 10 hours per day we will achieve 16,000,000 words per year. And this amount to around 50,000,000 words within 3 years. It should be pointed out that the process of selecting the texts would extend over the three years of the project.
- Encoding the texts with XML mark-up. Texts with different formats (Doc, PDF, HTML) will be converted into a unified framework (XML format) in which the texts will be enriched with features such as paragraphing and header information regarding text type, author, target audience, etc.
- Organizing procedure for annotating and proofreading the texts. Tasks include deleting extra and unnecessary material from texts and checking and adjusting paragraphing markers. This will be done manually.
- Obtaining letters of copyright. This will involve in the first place identifying the owners of web sources and finding the right addresses.

Stage 2:

- Tagging the corpus: Will use an automatic tagger, which is developed, in Lancaster University specifically for Standard Arabic.
- Assessing the performance of the tagger. This can be done, by running the tagger on a small corpus to detect unacceptable tags. The result can then be checked both manually and also by consulting another linguist researcher.
- Checking the applicability of the tagset to the data at hand and modifying it in order to give a better result for our corpus. Since the tagger has been developed for Standard Arabic we anticipate some problems when applied to our corpus.
- Investigating tools used in the British National Corpus and adopting them for use in the Arabic corpus.
- Validation and refinement of an existing XML system such as the one used for the BNC and adjust where necessary to suit Arabic.

Stage 3:

- Designing some tools that would be useful to use for processing the texts. They are such as: database that could be used as a search tool to recall texts of certain specifications, concordance program that could be used to process the texts
- Consulting teachers about the use of a corpus for teaching and about tools necessary to use with the corpus. Teachers might provide some new ideas of linguistic analyses to use, which we have not thought of before.

Commercial partners:

We are seeking some support from commercial partners such as publishers either here in Europe or in the Arab world. The intention is to get some help from them in developing the corpus by providing us with useful texts. This is useful for developing the base level of our corpus. In return, these partners can use the corpus for their benefit to develop some products such as specialized dictionaries. At present there are some general dictionaries such as Hans Wehr English-Arabic dictionary of Modern Written Arabic, Oxford English-Arabic dictionary of current usage and Al-Mawrid. There might be some interest in producing dictionaries for scientific terms or other specialized subjects. Another idea is developing a comprehensive lexicon that could be used as a source for teaching materials and writing exams.

Resources:

- **Staff:** the development of the project will require the employment of (a) an Arabic linguist as a post-doctoral research fellow for three years. Dr Latifa Al-Sulaiti is keen to take this post, as it will be a continuation of her MSc project. During her degree she had learnt some techniques and acquired some knowledge for developing her corpus, which she can use and develop further in the present project. (b) A computer scientist, preferably with knowledge of Arabic. The researcher we have in mind is Dr Shereen Khoja who is at present an assistant professor in computer science at Pacific University, USA. She has recently obtained her PhD for developing a part-of-speech tagger for Arabic. (c) Some research assistants who will help in proofreading of texts. These will require hourly pays.
- **Equipment:** in order to make best use of our corpus some language processing tools must be available. Our research software consists of two principal components: text processing tools and text analysis tools. Thus for the former there will be a need for a lemmatiser, a tagger, and an online dictionary. As for the latter we need a concordance program that generates a frequency list of words, or find relations among selected words, a database program that works as a search engine that searches for a text of specific nature or an author. As far as possible we are going to utilize the existing software resources with the intention of adopting them to suit the special nature of Arabic. Other equipment includes a powerful PC, which costs around £5000.
- **Travel:** our results will be reported on a regular basis and will be published in conferences and journals. This is important, as it will make the corpus more widely known. Thus there is a need for a reasonable travel budget. Some relevant conferences are such as Corpus Linguistics Conference (CL) and Teaching and Language Corpora Conference (TALC), which can be held in the UK or Europe.
- **Other Expenses:** it is essential that we get technical help from staff in the school of computing. Therefore, a sum of around £10,000 would be suitable to cover up this resource. There is also a requirement for a membership fee in the LDC for three years. This will give us an access to Arabic data that are available there. In addition, some funding is required for organizing at least two workshops to commercial partners. The first workshop is to explain our plan of the corpus and the second workshop will be to update of the progress of the work.

References:

Corpus Encoding Standard. <http://www.cs.vassar.edu/XCES>).

Beesley, K. (2001). Finite-State morphological analysis and generation of Arabic at Xerox research: status and plans in 2001. In <http://www.elsnet.org/acl2001-arabic.html>

Hoogland, J. (1996). The use of OCR software for Arabic in order to create a text corpus of Modern Standard Arabic for lexicographic purposes. *Proceedings of the international conference and exhibition on multi-lingual computing*, pp.2.7.1-2.7.16.

Ide, N. and Macleod, C. (2001). The American National Corpus: A Standardized Resource of American English. *Proceedings of Corpus Linguistics 2001*, Lancaster UK.

Khoja, S. et al. (2003). A tagset for the morphosyntactic tagging for Arabic. In Wilson, A, Rayson, P, and McEnery, T., eds. *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Lincom-Europa, Munich, pp.59-72.

Maamouri, M. and Cieri, C. (2002). Resources for Arabic Natural Language Processing at the linguistic Data Consortium, In *Proceedings of the International Symposium on: The Processing of Arabic*, Tunisia , pp.125-146.

Zemanek, P. (2001). Clara (Corpus Linguae Arabicae): An Overview. [Online]. Available from the World Wide Web: <http://www.elsnet.org/acl2001-arabic.html>.