

The design of a corpus of Contemporary Arabic

Latifa Al-Sulaiti and Eric Atwell
University of Leeds

Corpora are an important resource for both teaching and research. Arabic lacks sufficient resources in this field, so a research project has been designed to compile a corpus, which represents the state of the Arabic language at the present time and the needs of end-users. This report presents the result of a survey of the needs of teachers of Arabic as a foreign language (TAFL) and language engineers. The survey shows that a wide range of text types should be included in the corpus. Overall, our survey confirms our view that existing corpora are too narrowly limited in source-type and genre, and that there is a need for a freely-accessible corpus of contemporary Arabic covering a broad range of text-types. We have collected and published an initial version of the *Corpus of Contemporary Arabic* (CCA) to meet these design issues. The CCA is freely downloadable via WWW from <http://www.comp.leeds.ac.uk/latifa>.

Keywords: corpus, contemporary, Arabic, design, language variation, teaching Arabic as a foreign language (TAFL), Language Engineering

1. Introduction

Corpus building has grown widely in recent years. Corpora of various types have been developed for different research and teaching purposes (McEnery & Wilson 1996). English, being an international language, has received the greatest attention among the research community. There are several types of corpora that have been built, not only to investigate the main varieties of English, British (Johansson et al 1986; Aston & Burnard 1998) and American (Francis & Kučera 1979; Ide 2003) but other varieties such as Australian (Ahmad & Corbett 1987; Peters 1987; Green & Peters 1991), Indian (Shastri 1988), Cameroonian (Tiomajou 1993), and others. These corpora contain representative

samples of English text; many are also enriched or tagged with additional linguistic analyses, such as part-of-speech tags (PoS-tags) on each word showing its grammatical category or function in context (Leech et al 1983). English language corpora have been used in development of English language teaching materials, as well as language processing systems such as speech recognisers, spelling and grammar checkers, dialogue systems etc. (Atwell 1999).

Arabic is also an international language, rivalling English in number of mother-tongue speakers (Graddol 1997). However, little attention has been devoted to Arabic. Although there has been some effort in Europe and U.S.A. which has resulted in the successful production of some 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 lack of efficient tools such as tokenisers, taggers, morphological analyzers and optical character readers, which are necessary for developing and exploiting an Arabic corpus (Atwell et al 2004). Nowadays corpora, and bilingual parallel corpora in particular, are established tools in language learning and teaching, such as Teaching English as a foreign language (TEFL); and in Language Engineering (LE) and Machine Translation (MT) research and development. This paper discusses the development of a corpus of contemporary Arabic; the focus is on Arabic used in modern commercial and social communication, which should make this a useful resource for development and evaluation of Arabic LE as well as TAFL materials. Thus designing and compiling the corpus would depend on the views of these potential users and what they think would be effective for their needs.

2. Justification for a new Arabic corpus

Arab and European scholars who are interested in studying Arabic have developed several corpora, which can be an important research resource since contemporary Arabic needs some solid empirical linguistic 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. Table 1 (based on Al-Sulaiti 2004a) gives a brief description of existing Arabic corpora; these are not linguistically tagged (e.g. with PoS-tags), and are restricted to a specific source-type.

Table 1. Classification of Arabic Untagged Corpora

Name of Corpus	Source	Medium	Size	Purpose	Material
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	3 million 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	80 million 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 2 million 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	50 million 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	10 million words	Lexicography, general research, NLP	Unknown
An-Nahar Corpus (2001)	ELRA	Written	140 million words	General research	An-Nahar newspaper (Lebanon)
Al-Hayat Corpus (2002)	ELRA	Written	18.6 million words	Language Engineering and Information Retrieval	Al-Hayat newspaper (Lebanon)
Arabic Gigaword (2002)	University of Pennsylvania LDC	Written	Around 400 million	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	3 million words	Teaching translation & lexicography	Publications from Kuwait National Council

Table 1. (continued)

Name of Corpus	Source	Medium	Size	Purpose	Material
General Scientific Arabic Corpus (2004)	UMIST, UK	Written	1.6 million words	Investigating Arabic compounds	http://www.kisr.edu.kw/science
Classical Arabic Corpus (CAC) (2004)	UMIST, UK	Written	5 million words	Lexical analysis research	www.muhammadith.org and www.alwaraq.com
Multilingual Corpus 2004	UMIST, UK	Written	11.5 words (Arabic 2.5 million)	Translation	IT-specialized websites, computer system and online software help, one book
SOTETEL Corpus	SOTETEL-IT, Tunisia	Written	8 million words	Lexicography	Literature, academic and journalistic material
DARPA Babylon Levantine Arabic Speech and Transcripts (2005)	University of Pennsylvania LDC	Spoken	About 2000 telephone calls	Machine translation, speech recognition, & spoken dialogue system	Fisher style telephone speech collection

The above corpora 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 (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, who do not have institutional membership of LDC or ELRA. 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.¹ There are even free internet-based services which allow teachers and researchers to PoS-tag their own English corpus texts automatically; Atwell et al (2000b) report that this opened up English corpus resources to a much wider audience, for example English language teachers used the PoS-tagger service to set online classroom exercises on English grammar.

The corpora in Table 1 represent raw material except the *General Scientific Arabic Corpus* of which 1 million 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 2003). However, this corpus is new and not (yet) in the public domain; furthermore, this size of 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.

2.1 The usefulness of an Arabic Corpus

Since Arabic texts are becoming widely available on the web, and some linguists advocate use of Web data for research (e.g. Bergh 2005), one may ask the question: do we really need a corpus?

The purpose of a computer corpus is not merely to gather a big file of different texts and store it on the computer, but rather to prepare the texts and put them in a certain format so that they can be used by search tools and the results of the search can be displayed in a way that is meaningful and useful to the linguist, teacher and learner especially at the advanced level. For example, teachers and learners can explore the use of a word in different types of texts to see how frequently this word is used, how many meanings it has, what syntactic environment it occurs in, whether the word has the same frequency of occurrence in all types of texts. Teachers can identify the most frequent words and select them as a basis for their material. There is also the study of syntactic structures and analysing the distribution of competing structures. For example, the uses of verb–subject vs. subject–verb word order in Arabic: which word order is more preferred in children’s stories, interviews, and scientific documents? Also the uses of passive and active forms: is passive in Arabic used more often in scientific than literary writing? In addition, it is interesting to compare the language of the written and spoken at the level of lexis as well as the level of syntax. The corpus is to be annotated with XML mark-up which includes information about the text, author, and source; this gives the opportunity to conduct empirical analyses which control extra-linguistic factors (such as age, sex, region, social class, and education level) and examine the accompanying linguistic variations. We hope our corpus would be further enriched with other information such as tagging which signifies information on word classes. This would make retrieval of useful information qualitatively and quantitatively much richer and easier to handle.

A limited amount of linguistic research has been carried out using existing Arabic corpora; very limited when compared with corpus-based English linguistics. Parkinson and Farwanah (2003) note that corpus-based research has only featured recently in the Annual Symposium on Arabic Linguistics: “four of the papers deal with the area of corpus linguistics (new for this series), including papers from both a computational and a variationist point of view”. The four papers used corpus-based evidence in studies of Arabic particles (van Mol 2003a; Parkinson 2003), noun-phrase structure types (Al-Ansary 2003), and lexical productions (Taylor 2003).

However, interest in corpus-based empirical research is on the increase in Arabic linguistics. The following examples illustrate a range of avenues for further linguistic research given richer Arabic corpus resources.

One of the first corpus-based studies of Arabic was Parkinson's (1985) sociolinguistic investigation of terms of address system in Egyptian Arabic. The aim of this study was to define who is using what terms, to whom and in what situations. In order to achieve this goal, it was essential to gather a large quantity of natural data. Parkinson states:

It was felt that survey and interview techniques, while proving valuable information, were not an adequate substitute for a large dose of the actual raw data of naturally occurring speech. (Parkinson 1985:4)

Thus over a period of one year he gathered data from several parts of the city of Cairo to get representative examples covering varied social class and life style. He collected over five thousand instances of terms of address which represent a large number of social situations. This natural corpus which forms the basis of the investigation was processed manually. Each instance of term of address was recorded on a card with information about the setting, sex, age, social class etc. The data was then analysed from both quantitative and qualitative points of view.

Recently in the area of general linguistic research, van Mol (2000b) recognised the importance of using corpora. Although it is more complex to explore Arabic corpora due to its polysemic nature, lack of vocalisation (short vowels), and the affixation of function words to content words, he demonstrated that using corpora for linguistic analysis would give a new insight into the structure of Arabic. A tagged corpus of Arabic news broadcasts (240,000 words) from the radio station of Algeria, Egypt, and Saudi Arabia was compiled. The corpus showed that the future particles 'sa' and 'sawfa' do not have any distinction between their use in the near and remote future as it is assumed traditionally. The corpus also showed regional variation in the use of Modern Standard Arabic. For example, the expressions *fi nafs il-waqt* and *fi-lwaqt nafsihi* 'at the same time' occur with the same frequency but when looking at their use in the different countries, it was found that *fi-lwaqt nafsihi* occurs most intensively in the Egyptian corpus, while it is missing from the Algerian corpus. But in the Saudi corpus the two expressions have the same distribution. A more detailed corpus study on the complementary particles in the three above dialects was also pursued (van Mol 2003a, 2003b).

Gully (1997) made an analysis of the language of advertising in Arabic. He investigated the discourse of commercial consumers advertising in the written

and visual media of Egypt. The study is based on a corpus of approximately 150 newspaper and magazine advertisement and television commercials. He came to the conclusion that advertisement reflects the culture and its people and it also mixes linguistic levels or codes.

Al-Muhanna (2003) investigated new compounds in Arabic scientific writing in a corpus of 1 million words. The result was compared with terms used by the language academies. Further linguistic analyses on written and spoken Arabic have been also made (Al-Ansary 2003; Parkinson 2003; Taylor 2003).

In the area of lexicography, corpora proved to be an invaluable source for developing good dictionaries. Using a corpus would enrich the dictionary with idiomatic expressions, illustrative phrases and collocations. Most existing dictionaries of Modern Standard Arabic are not corpus based. Hence, they are characterised by giving literal translations of a word, they lack collocations, phrasal verbs, and new words entering the language. Ghazali and Braham (2001) commented that:

Arabic dictionaries are notorious for representing a fossilised version of our language as each one is to a great extent a reflex of the preceding ones in content, (the same examples often taken from the Koran) and the way different meanings are ordered. This state of affairs is not likely to help the great majority of Arabic learners whether Arab school children or speakers of other languages. (Ghazali & Braham 2000:7).

Based on an Arabic corpus of 1,500,000 words they investigated the meaning of the verb *axadha* 'he took'. They found that in addition to its literal meaning there are two other meanings: 'to start doing something' and 'to take into consideration' when followed by certain constructions. The latter special sense provided 11% of occurrences; yet neither the standard traditional dictionary *Al-Wasiit* (1960) nor other more recent monolingual dictionaries has mentioned this meaning.

Researchers such as van Mol (2000a), van Mol and Paulussen (2001), Hoogland (1996) and Zemanek (2001) used corpora for developing dictionaries for Modern Standard Arabic. Van Mol developed a relational database *Ara-Lat* to build a Dutch–Arabic/Arabic–Dutch dictionary. The database contains a learners' dictionary which consists of 19,000 Arabic words translated in context and more than 10,000 illustrative sentences selected from a representative corpus of 3,000,000 words. Hoogland developed a corpus of over 2 million words for developing Dutch–Arabic/Arabic–Dutch dictionaries. Zemanek developed an Arabic–Czech dictionary which is based on a corpus of 50 million words.

Another application for corpora is computational linguistics and the development of speech and language technology. For example, Messaoudi, Lamel and Gauvain (2004) investigated transcription of Modern Standard Arabic broadcast news data using speech recognition. A corpus of 50 hours of audio data from 7 television and radio sources and 200 million words of newspaper texts were used to train the acoustic and language models. Language engineering researchers are aware of the need for improved Arabic corpus resources. For example, Xu, Fraser and Weischedel (2002) explored a number of strategies for retrieving Arabic documents using the TREC Arabic test dataset. Since they found their strategies improve retrieval performance, they want to validate them on more corpora when they are available. Their view is that ‘...the lack of a realistically large test corpus has been a problem in past studies on Arabic retrieval.’ (2002:269). Guidère (2002) discusses the use of Arabic corpora in machine translation systems development, but notes: ‘In the current state of machine translation research from and into Arabic, no reference corpus is yet available’ (Guidère 2002:1).

3. The composition of existing corpora

In this section we will discuss the basic composition of the corpora that deal mainly with general Modern Standard Arabic. We must point out that there is no formal classification of the genres and there is no exact figure for the number of words in each genre in most of the corpora we have listed in our table. The only exception is CLARA. We obtained the information through contacting the corpora’s developers. Some of them replied that they have not investigated the genres or their size as the main focus is to use the corpus for their own specific research rather than making it available for general use.

In contrast, the developers of Brown (Francis & Kučera 1979) and LOB (Johansson et al 1978, 1986) corpora of American and British English aimed for a wider range of genres to satisfy a wider range of potential users: each million-word corpus contains 500 text samples of approximately 2,000 words distributed over 15 text categories, as shown in Table 4.

This model was followed by several subsequent English corpora including FLOB, FROWN, ACE and the Wellington NZC. However, this Brown/LOB composition was designed at a time when very little text (let alone speech) was available in computer-readable form, so it focused on published, printed text sources. Later projects extended sources to include transcriptions of spoken English, and various informal, unpublished sources of text. The *International*

Table 2. Text categories in some of the existing corpora

Corpus	Text categories
Nijmegen Corpus (from Hoogland p.c. 2004)	Literary texts Popular scientific Magazine articles Newspaper articles Press agency items
LDC Corpus (from Maamouri p.c. 2004)	Newswire (limited to political news / sometimes there are news about sports and economy) Broadcast Telephone conversation which include inter-dialectal variation
Al-Hayat Corpus (from ELRA homepage 2003)	General Car Computer News Economics Science Sport
An-Nahar Corpus (from An-Nahar homepage 2004)	General Politics News Computer Literature Economics Sport
E-A parallel Corpus (from Al-Ajmi 2003)	History Economy Arts Literature General science

Corpus of English (ICE 2003) is an international project involving a consortium of 16 research groups, each collecting one million words of English as spoken in their own country, following an agreed set of procedures including the distribution of categories shown in Table 5 (Greenbaum 1996).

The *British National Corpus* developers aimed for a much larger corpus, a hundred million words; and they started at a time when many more sources of machine-readable language were available; so the BNC adopted a more complex system of categories. According to (BNC 2003): "...The Corpus is designed

Table 3. Text categories and number of words in CLARA (from Zemanek 2001)

Text categories		Number of words
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

Table 4. Text categories in English Brown and LOB corpora (from Johansson et al 1986).

Text categories	Number of samples in each category	
	Brown Corpus	LOB Corpus
A Press: reportage	44	44
B Press: editorial	27	27
C Press: reviews	17	17
D Religion	17	17
E Skills, trades and hobbies	36	38
F Popular lore	48	44
G Belles lettres, biography, essays	75	77
H Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30
J Learned and scientific writings	80	80
K General fiction	29	29
L Mystery and detective fiction	24	24
M Science fiction	6	6
N Adventure and western fiction	29	29
P Romance and love story	29	29
R Humour	9	9
Total	500	500

Table 5. Text categories in ICE: International Corpus of English corpora (from Greenbaum 1996); numbers in brackets show the number of 2,000-word texts in each category.

Spoken (300)	Dialogues (180)	Private (100)	Conversations (90) Phonecalls (10)
		Public (80)	Class Lessons (20) Broadcast Discussions (20) Broadcast Interviews (10) Parliamentary Debates (10) Cross-examinations (10) Business Transactions (10)
	Monologues (120)	Unscripted (70)	Commentaries (20) Unscripted Speeches (30) Demonstrations (10) Legal Presentations (10)
		Scripted (50)	Broadcast News (20) Broadcast Talks (20) Non-broadcast Talks (10)
Written (200)	Non-printed (50)	Student Writing (20)	Student Essays (10) Exam Scripts (10)
		Letters (30)	Social Letters (15) Business Letters (15)
	Printed (150)	Academic (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
			Popular (40)
		Reportage (20)	Press reports (20)
	Instructional (20)	Administrative Writing (10) Skills/hobbies (10)	
	Persuasive (10)	Editorials (10)	
Creative (20)	Novels (20)		

to represent as wide a range of modern British English as possible. The written part (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part

(10%) includes a large amount of unscripted informal conversation, recorded by volunteers selected from different age, region and social classes in a demographically balanced way, together with spoken language collected in all kinds of different contexts, ranging from formal business or government meetings to radio shows and phone-ins. The written texts were selected for inclusion in the corpus according to three independent *selection criteria*: domain, time, and medium. Target proportions were defined for each of these criteria. The *domain* of a text indicates the kind of writing it contains. 75% of the written texts were to be chosen from *informative* writings: of which roughly equal quantities should be chosen from the fields of applied sciences, arts, belief & thought, commerce & finance, leisure, natural & pure science, social science, world affairs; 25% of the written texts were to be *imaginative*, that is, literary and creative works. The *medium* of a text indicates the kind of publication in which it occurs: 60% of written texts were to be books, 25% were to be periodicals (newspapers etc.), between 5 and 10% should come from other kinds of miscellaneous published material (brochures, advertising leaflets, etc), between 5 and 10% should come from unpublished written material such as personal letters and diaries, essays and memoranda, etc; a small amount (less than 5%) should come from material written to be spoken (for example, political speeches, play texts, broadcast scripts, etc). The time criterion refers to the date of publication of a text: being a *synchronic* corpus, the BNC should contain texts from roughly the same period.”

The *American National Corpus* represents a fourth alternative approach to corpus design. In contrast to Brown/LOB, ICE, and BNC, aimed to fill a pre-defined set of categories in specific proportions, the ANC developers aim to include texts as long as they are current, available, and American, to simplify compilation of a large corpus on a limited budget and timescale; Table 6 (from ANC 2003) shows the composition of the ANC first release.

We have already argued that existing Arabic corpora are too restricted in source-type, so our CCA design should learn from Brown/LOB, ICE and BNC projects in including a wider range of text types and genres. However, given our limited budget and time, the ANC pragmatic approach is also attractive: we could aim to include texts which are current, available, and Arabic. As a compromise, we decided to poll potential users on the range of text-types to aim to include, but not to rigidly proscribe the proportions of each text type.

Although there is a growing interest in developing Arabic corpora for research, existing corpora are restricted in use. They were generally collected for a specific research project rather than as a resource for TAFL practitioners, and are not representative of Contemporary Arabic, and/or are not available to

Table 6. Contents of the ANC First Release

Text type	Text name	No. of texts	No. of words	Contributor
Spoken	Callhome	24	50,494	LDC
Spoken	Switchboard	2320	3,056,062	LDC
Spoken	Charlotte Narrative	95	117,832	Project MORE
TOTAL SPOKEN			3,224,388	
Written	New York Times	4148	3,207,272	LDC
Written	Berlitz Travel Guides	101	514,021	Langensheidt Publishers
Written	Slate Magazine	4694	4,338,498	Microsoft
Written	Various non-fiction	27	224,037	Oxford University Press
TOTAL WRITTEN			8,283,828	
TOTAL CORPUS SIZE			11,508,216	

others because of copyright and other legal issues. Some corpora have recently become available for purchase from LDC or ELRA; but since teachers are not yet aware of the importance of corpora for teaching, not many institutions or individuals are ready to pay for a corpus. Stevens (1993) divides teachers into three categories: those who advocate the use of corpora, those who have never heard of them and those who are not sure of their benefit for teaching. Although more and more English language teachers fall into the first category, the use of corpora for teaching Arabic is extremely limited. The English department in the University of Kuwait is the only place we know of which regularly uses corpora for teaching Arabic translation and lexicography (Al-Ajmi 2003). Therefore, our aim is to consult TAFL practitioners on the design of a corpus of contemporary Arabic which will match their preferences, and to make our corpus available for public use via the WWW. As for copyright, we are contacting owners of Arabic source texts and getting agreement from them. So far we have received encouraging replies and most of them stress the condition that we should not charge users and learners for the use of the corpus. As far as the text types found in newspapers such as *An-Nahar* and *Al-Hayat*, it is true that they contain a wide range of texts. It has been recognized (Rademann 1998) that newspapers are a popular resource for building a corpus and that big parts of major corpora such as the BNC and LOB depend heavily on newspapers articles. But the above corpora are restricted to one regional environment (Lebanon). If there is any occurrence of some regional expressions they would be restricted to that particular region only. In addition, some text types in newspapers such as science are not handled with sufficient depth.

4. 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 (MSA) (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 to foreign learners (Ferguson 1971). But for the past twenty years or so spoken Arabic has become as important as Standard Arabic especially since MSA does not provide a realistic means of everyday communication. Thus there has been a debate over which dialect should be taught, and whether it should be taught before the Standard Arabic or after it. There are some who support teaching the standard before the '*ammiyya*, while others support teaching both the standard and the '*ammiyya* at the same time (Younes 1990). Still others support the teaching of ESA or *al-lugha al-wusta* before the Standard (Nicola 1990), or *al-lugha al-wusta* after Standard (Haddad 1985). There is also variation in the regional or national varieties to focus on; for example, a survey conducted by Elkhafai (2001) found out 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 or *al-lugha al-wusta*, 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' (Holes 1990:37).

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

So, we can define the term 'Contemporary Arabic' by the form of Standard Arabic 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. Contemporary users of Arabic can naturally "code-switch", for example speech or writing on religious topics may include extracts or quotations from the Qu'ran or other classical sources; this is more natural in contemporary Arabic than, say, contemporary English, so we do not see the need to enforce a rigid bottom time limit for extracts or quotations in texts.

In teaching foreign learners it is practical to choose one specific variety along with the Standard. Students get to 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 programs such as the one at the University of Leeds. However, although native speakers should understand the learner, this does not guarantee that the learner would understand Arabs who speak other varieties. Even though the Egyptian dialect is traditionally known to be the most understood dialect by other speakers, this does not mean that Egyptians themselves understand speakers of other varieties especially if they have little 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. Highly educated people are equipped with knowledge to modify their speech to ease communication with foreign speakers but most ordinary people are not.

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 (Thomson 1994:18).

With the growth of satellite broadcasting, listeners and viewers have access to many different channels and thus different spoken varieties. For example, the Lebanese channels LBC and Future have a wider audience than many other Arabic channels. This means that the Egyptian dialect may soon no longer be the dominating variety.

So, the plan of our corpus is 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. Our main focus is to represent the standard form, written and spoken, as well as some regional varieties, for example as reflected in a range of Arabic broadcast media. The corpus will be a rich resource for learners to explore, compare and learn about the present Standard Arabic with its new vocabulary and its different regional varieties.

In the inaugural issue of the Arabic Computing Society newsletter, Mili (2003) expressed his disappointment with the level to which Arabic has sunk. In the fields of science and technology, both teaching and research are conducted in English. He feared that Arabic may be approaching the level of extinction in science if there is no collective effort from the public and private sectors to revive it. Ironically, Mili's article, along with all the other articles in the Arabic Computing newsletter, was in English. The steps he suggests for reviving Arabic include teaching such subjects in Arabic and providing tools and resources to support the use of Arabic around the world. We hope our corpus, which would include some technical and scientific material as part of its content, would become one of the resources that contribute to the teaching and exploring of the language of science. In addition, this corpus would be the source we plan to utilise for developing Arabic language teaching materials (Al-Sulaiti & Knowles 2002).

5. Sources of corpus texts

Based on our survey and on the views expressed by linguists and computer scientists, we can conclude that there is a demand for developing a more balanced Arabic corpus that would include texts other than newspaper documents. This corpus should be freely available for the public via the WWW, particularly for teaching Arabic as a foreign-language, and for use in both language processing

and general linguistics research. Our first step in pursuing this aim is to make a decision on the type of texts to be included in the corpus. Some corpora developed after the Brown and the *British National Corpus* (BNC) seem to emulate their styles regardless of the suitability of the sources or categorizations of the topics. However, for this corpus we should make our decision based on the needs and views of end-users: language teachers, and language engineers developing software systems for language processing (Atwell 1999).

Due to limitation of budget and time, our initial target was to compile a corpus of 1 million words, comparable in size to LOB, Brown, ICE-GB etc. This may seem small compared to BNC and ANC, but a 1-million-word corpus is still a potentially useful resource. Although Brown and LOB were collected about 30 years ago, they are still used in current research (e.g. Leech & Smith 2005; Wilson 2005). Similarly, 1-million-word corpora have proven useful for comparative studies in the ICE project; and corpus linguistics researchers have used even smaller corpora in studies of, for example, learner error-types (Atwell et al 2003), grammar checkers (Atwell 1987), language technology evaluation (Atwell et al 2000b; Elliott et al 2003, 2004), specialist language in science (Dury 2004), variation in use of contractions (Peitsara 2004), attribution markers (Murphy 2005), grammatical inference (Atwell 1988), language visualisation or animation (Abu Shawar & Atwell 2005). Our data would be wholly derived from texts received in machine-readable form. Capture of handwritten or spoken data via optical scanning or audio-typing transcription would have to be avoided as this has practical limitations and is a very slow and/or expensive process.

5.1 Written data

Nowadays in most Arab countries, publishing companies produce large amounts of material on the World Wide Web. Thus there are a growing number of texts available in machine-readable form and we have identified over 40 promising sources, containing texts on a wide range of topics including short stories and children's stories. These are the genres that are generally reported to have the fewest texts on the web. Below is a list of the sites who have granted permissions to use samples of their texts in our corpus (at the time of writing; we are still expecting more replies to come)

1. Majallat al-Arabi: <http://www.alarabimag.com> (Kuwait)
2. Majallat Ofouq: <http://www.ofouq.com> (Saudi Arabia)
3. Al-qissa Al-Arabiya site: <http://www.arabicstory.net>

4. Majallat PC Al-Arabiyya: <http://www.pcmag-arabic.com> (UAE)
5. Arabic BBC site: <http://www.BBCArabic.com> (UK)
6. Majallat Sayyidaty: <http://www.sayidaty.net> (UK)
7. Majallat 'aalam Al-'iqtisaad: <http://www.ecoworld-mag.com> (Saudi Arabia)
8. Majallat Nizwa: <http://www.nizwa.com/> (Oman)
9. Al-Dawriyya Al-Tibbiyya Al-Arabiyya: <http://arabmedmag.com> (Syria)
10. Aklaat site: <http://aklaat.com/> (UAE)
11. Islam on line site: <http://www.islamonline.net> (Qatar)
12. Al-Raay Al-'aam newspaper: <http://www.alraialaam.com> (Kuwait)
13. Majallat Al-Ma'rifa: <http://www.almarefah.com> (Saudi Arabia)
14. Majallat Akhir Saa'a: <http://www.akhbarelyom.org/akhersaa> (Egypt)
15. Al-Kumpyuter fi Al-'aalam Al-Arabi: <http://www.arabcomputing.com> (UK)
16. 'aalam Al-kumpyuter: <http://www.alamalcomputer.com> (Egypt)
17. Al-Raya Newspaper: <http://www.raya.com> (Qatar)
18. 'uluum wa tuknologia <http://www.kisr.edu.kw/science/> (Kuwait)

As far as our sampling principles are concerned our criteria would be based on the results of our survey and the needs of end users. The text categories which score high in our questionnaire should have the biggest number of words. In major corpora samples are specified to contain 2000 to 5000 words or even 20,000 words (Oostdijk 1988) to give more reliable results. The problem with using the WWW as the source of material is that it is hard to find samples of this size. Most of the material is short articles or a group of articles that deals with a specific topic. It is very rare to find a web-page the length of a book on the WWW.

Because of this disadvantage we tried to make use of all these available sources and include as many texts as we could find. Therefore, the size of the texts in each genre was not limited to a specific number of words as is the case in most corpora. In this regard, our corpus follows the design of the American National Corpus (ANC) (Ide 2003).

5.2 Spoken data

We believe that having a reasonably large representation of spoken recordings is an important part of our ideal design for the corpus, even though practical limitations prevent us from including much spoken text in the first version of

the corpus. Our longer-term aim is to work with some Arabic broadcast media services and obtain spoken recordings covering topics such as general speeches, news, interviews, plays, narrations, poetry reciting, phone-ins and other programs that might be useful. However, we might exclude speeches such as preaching as this kind of genre contains Classical Arabic. We would hope to receive recordings on audio cassettes or on CD's. We would have to first digitize the cassette recordings using a program such as CoolEdit which allows us to convert the original sound files to WAV files; then we would have to deal with the transcription and producing electronic copies. Clearly, it would be easier to try to use WAV files on websites where we have copyright permission, particularly on some websites where broadcast transcripts may be available.

However, even in the best case, capture of spoken data is much more labour-intensive (and thus expensive) than capture of written text from websites; so in our 1-year project, the initial version of our *Corpus of Contemporary Arabic* will not include spoken texts, but will have a "place-holder" waiting to be filled.

6. User survey: Methodology

Our choice of text types should reflect the needs of users. So, we carried out a survey of language teachers and language engineers to get their opinions on the texts that might be of use to them (Al-Sulaiti & Atwell 2003). We developed an online questionnaire and made it available via mailing lists for language teachers and language engineers.² We also sent it to some individual teachers. The questionnaire consisted of three sections. Section 1 contained personal detail questions covering the name of their institution or 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). 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 multilingual corpus for machine translation evaluation (Elliot et al 2003). More types were added from knowledge of other corpora. In order to obtain a balanced corpus, our selection of the texts can be classified under two broad categories: texts that represent factual knowledge and texts that represent creative knowledge. In this regard we are following the guidelines used in other existing corpora such as the BNC. The texts belong to the following major categories:

Written: Fiction – Arts – Science – Business – Miscellaneous

Spoken: TV – Radio – Conversation

Section 3 contained one question for language engineers and 14 questions for language teachers. The purpose of these questions was to examine the factors which affected their choice of texts, and to get their views on any other text that could be added.

7. User survey: Results and discussion

We divided the respondents into the two groups: language teachers and language engineers. 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. We therefore had to calculate only two values: ‘useful’ against ‘not useful’. We calculated the responses of language teachers to show their most useful texts; then did the same 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.

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:

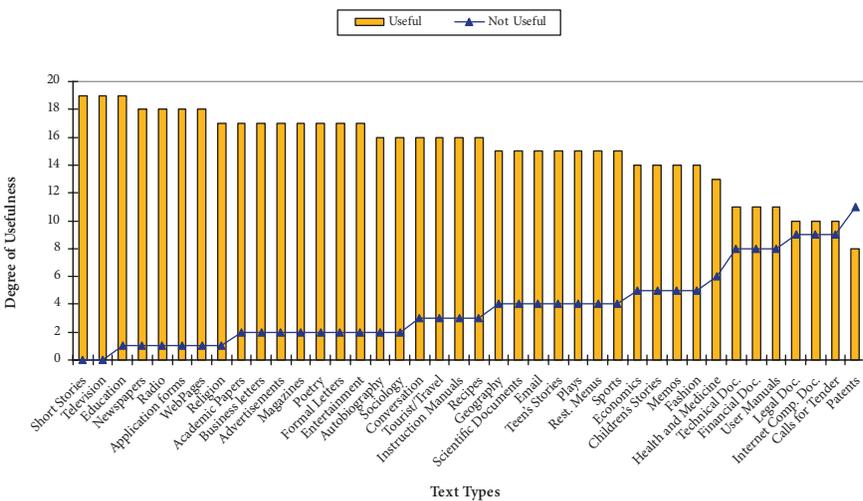


Figure 1. Distribution of the useful texts by language teachers

Category 1: 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.

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.

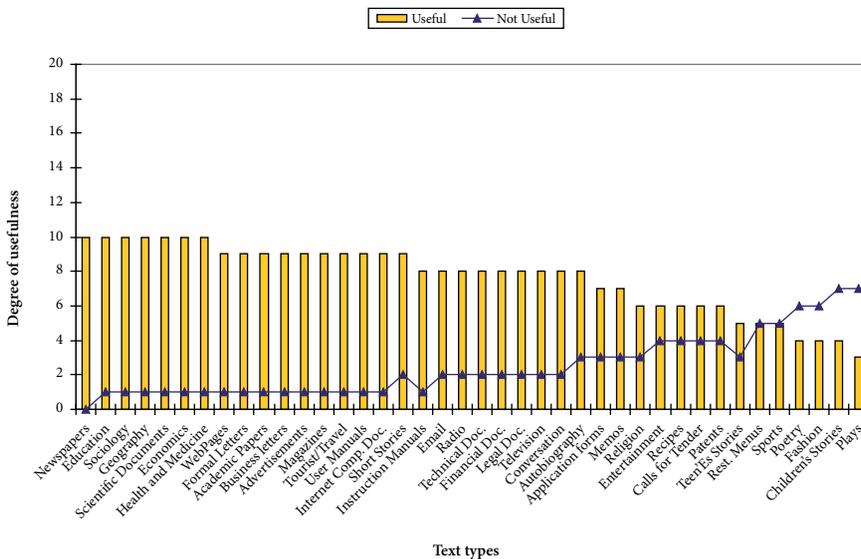


Figure 2. Distribution of the useful texts by language engineers

Category 1: 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 computing documents, short stories.

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

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

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 we find it surprising that academic subjects were classified at the top of the list.

From this result we are now able to make our selection of the texts that we think should occupy the major part of the corpus. The texts that have been marked as less useful in both groups would be included but with fewer words. Even the less useful categories were judged "useful" by some of the respondents, so we should not exclude these entirely. Overall, our survey confirms our view that existing corpora are too narrowly limited in genre; although the survey results do not proscribe exact proportions of each text-type, there is a need for a corpus of contemporary Arabic covering a broad range of text-types.

We will now discuss briefly the other parts of the questionnaire, which have some reflection in the design of the corpus. The questionnaire asked the users to identify the potential future applications of the corpus and give their own suggestions for any other applications.

The ten potential Language Engineering applications suggested in the questionnaire, in order of respondent preference, were:

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

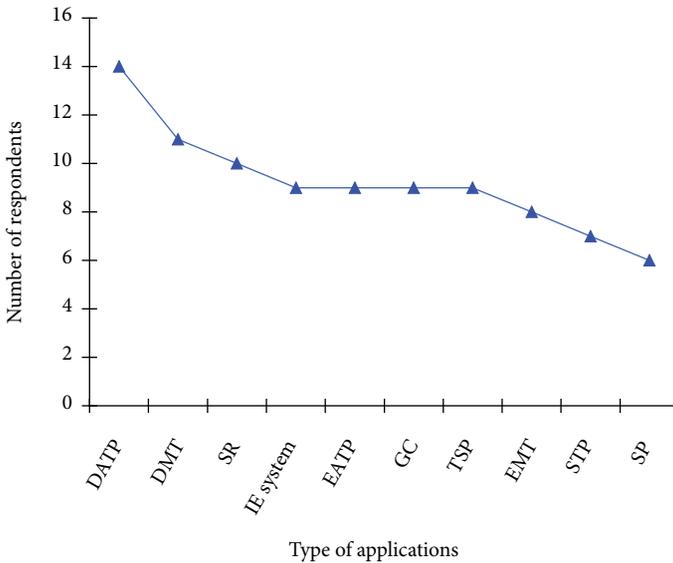


Figure 3. Number of responses for future applications of the corpus

- Speech to text processing (STP)
- Speech production (SP)

The preferred types of applications for the corpus that were identified are shown in Figure 3. The second highest score of potential application for the corpus is ‘developing MT’. This was interesting to us as we wanted to include parallel Arabic/English texts but we needed some justification or support from the users. In addition to the support for machine translation applications, one respondent suggested using the corpus for translation studies. This purpose cannot be achieved unless we include some parallel texts. 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 we received 3 suggestions by language engineers for including parallel texts. Based on the result in Figure 3 and on the opinions of some of these respondents we believe that including parallel texts in the corpus is 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.

Although the corpus might appear to be not large enough to be used at this stage for all the applications listed above, we demonstrated previously (Section 5) that for some research it is unnecessary to have a large corpus. For example, Elliott et al (2003, 2004) compiled a multilingual corpus for the purpose

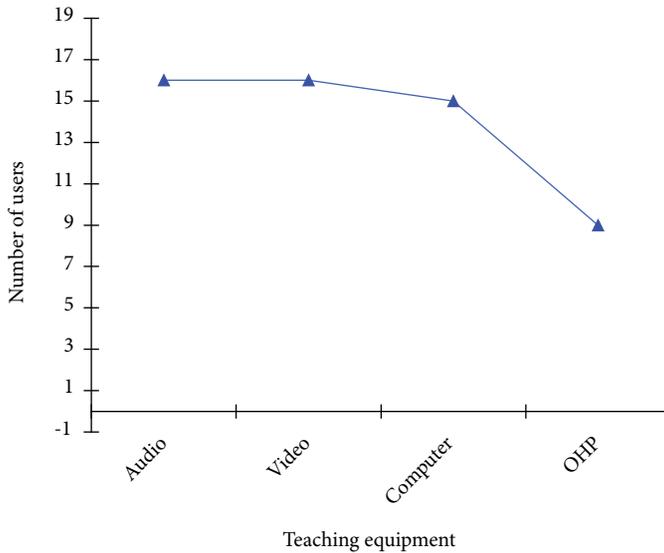


Figure 4. Survey in the use of teaching equipment

of evaluating the output from 4 MT systems. Her preliminary investigation confirmed that reliable scores can be obtained from using a small corpus; the corpus used consists of 40 source texts, each of which contains 400 words for each language pair (around 16,000 words in total). However, the text types must reflect the users' needs in that they cover a specified range of categories (software user manuals, technical press releases, frequently asked questions, technical reports, legislative documents, medical documents).

We asked another question about the teaching equipment available for Arabic. Our main purpose was to assess how much computers are used for teaching Arabic. If the result was 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 a decline in the use of the OHP. The survey shows that the computer is as widely-used in teaching as other equipment.

One of the important issues regarding the content of the corpus is that we are planning to include some written texts or spoken texts, which contain colloquial forms, as we believe such types of texts represent contemporary Arabic. One possible source is Internet chat sites, which are characterized by their informality. We are not sure if such texts are acceptable or useful; among the questions we asked was whether teachers approve of teaching registers to foreigners. The results we obtained from this question can be seen in Figure 5.

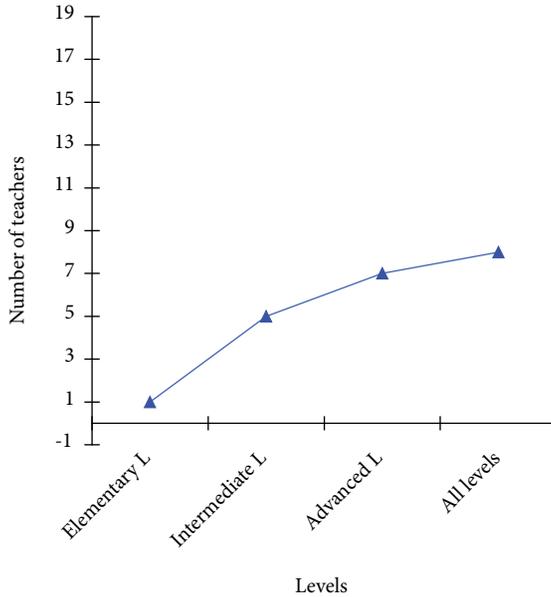


Figure 5. Number of teachers supporting teaching registers

The majority of users agreed that registers are useful for teaching foreigners. However, the highest score was for teaching it at advanced levels, or across all levels. This supports the inclusion of colloquial forms in the corpus. In support of our 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, programs that solely teach Modern Standard Arabic cause ‘greater difficulty of application to the local environment’ (Lunt 1992:122).

8. The initial Corpus of Contemporary Arabic

In compiling the CCA, its internal structure was designed 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 copyright permissions. Fortunately, most of those contacted were pleased to use their material for teaching. Once copyright was granted, text collection commenced. There are some search engines that can be used to “trawl” for a corpus such as WebCrawler; but in this instance, the texts were collected manually because only those sites who had given their consent were used. Every text was

encoded with a header with the necessary information added, and saved as an XML document.

At the end of our 1-year project, we published an initial version of the *Corpus of Contemporary Arabic* on the World Wide Web.³ This initial version of the CCA was presented at TALC'04 in Granada (Al-Sulaiti and Atwell 2004), and used to demonstrate portability of the XAIRA concordancer (Burnard 2004) to Arabic. So far the CCA consists of over 843,000 words in 416 files covering a wide range of categories. 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 web pages. Table 7 shows the text categories which are derived from any of the sources, the number of texts in each category, and number of words.

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 if we were to include a significant proportion of spoken source transcriptions and parallel Arabic-English translations, since these were not readily available, and it would have been prohibitively time-consuming to transcribe or translate texts ourselves. Rather than “fill up” the million-word target with printed texts, we decided to leave a gap for spoken and parallel

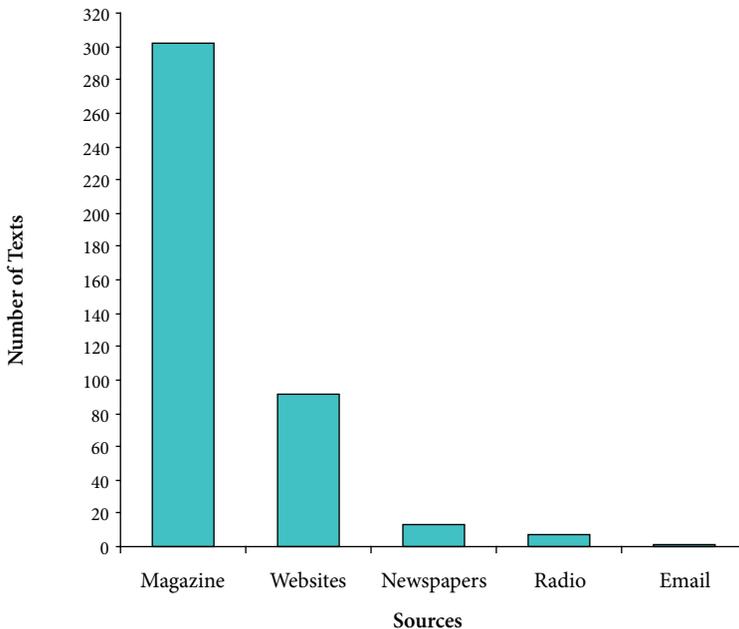


Figure 6. Number of texts used from the different sources

Table 7. Number of texts and number of words in each category

Text Categories		No. of texts	No. of words
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

Arabic-English texts, to be filled later in a follow-up project to extend the *Corpus of Contemporary Arabic* to Version 2.

Figure 6 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.

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

8.1 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 paper 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 2004:1745):

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

Sharoff applied his scheme to classification of texts in large corpora of English, Russian, German and Chinese compiled from Internet sources (Sharoff forthcoming). The classification of the texts of the CCA is based on Sharoff's as it is designed for internet-sourced corpora across a range of languages beyond English, and it seems to group a variety of text types under a general domain

Table 8. CCA text types classified in Sharoff's domains

Domains	Text types
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

which is quite tidy. Table 8 shows a rough classification of the text types included in the CCA within Sharoff's general domains.

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. Even though they were published in the 1960's or 1970's it was felt that works of such authors must be included for learners as they represent the best contemporary 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.

Texts for the religion category would obviously contain extracts or quotations of much older texts such as Hadith or Qur'an. But this does not in any way contradict the general aim of building a contemporary corpus. Having old words or expressions would be only a reflection of one specific genre rather than the corpus as a whole.

While collecting the texts we created a database which stores the ID number, title, source, number of words, year of publication, and author's name of each text. This database served as a corpus management tool for the organisation of the texts of the corpus, e.g. for counting the words automatically.

8.2 Number of texts

As can be seen from Table 6 above not all the text categories were obtained, 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*; many published children's books have text which is handwritten rather than word-processed, to make the text easier to read (Al-Sulaiti 1991). As for spoken recordings from radio and TV, only a small sample was obtained.

8.3 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 complete book could not be found. Texts that contained tables were avoided because when saving the file as a text the information in the table is hidden or lost. 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.

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 each with their own header. Also, this would create overheads for concordancing. Therefore, in such cases 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 cite the names of all the authors.

8.4 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 response delays with some sources such as science, Internet computing prevented this goal from being achieved. In addition, it was extremely difficult to complete this project within the one-year program limit. Despite this, a good number of authorisations were obtained for magazines, newspapers, and websites.

9. Conclusions and future plans

In this paper we provided an extensive survey of currently available Arabic corpora, mainly based on information on the Internet. Also we had to contact the people who are involved with this research to obtain some specific information or check the accuracy of information at hand. In so doing we have found concrete evidence for the need for a new Arabic corpus. We envisage that not only would this corpus fill a gap in the general field of corpus linguistics but it would also have a role in providing authentic material for teaching Arabic as a foreign language, developing tools that serve the spread of the use of Arabic, and encouraging wide scale research into investigating linguistic phenomena based on a large, varied dataset. The initial version of the *Corpus of Contemporary Arabic* is finished and available online at <http://www.comp.leeds.ac.uk/latifa/> and it is freely accessible for users, unlike other Arabic corpora.

As an extension to this project, we plan to set up infrastructure and prototype sampler corpus for the *International Corpus of Arabic*, an international collaborative research programme to parallel the *International Corpus of English*. An international steering panel of stakeholders will establish agreed standards for text types and categories; encoding and XML mark-up; morphological analysis, lemmatisation and part-of-speech tagging; and parallel English translations. We will provide a knowledge management environment for computer-supported collaborative work including discussion and authoring of standards, and tools for collation, mark-up, lexico-grammatical analysis, exploration and dissemination of the *International Corpus of Arabic*. We will collect, mark-up and lexico-grammatically annotate a 1-million-word sampler corpus as a representative subset of the full ICA; and use Arabic concordance and corpus exploration tools to analyse lexical and grammatical variation across the contributing dialects of Arabic in this sampler corpus. This will establish Britain at the centre of international development and exploitation of Arabic corpus linguistics and language engineering.

Acknowledgments

We would like to thank all those who participated in our corpus user survey; and all source owners who generously donated texts for inclusion in the online *Corpus of Contemporary Arabic*.

Notes

1. <http://thetis.bl.uk/lookup.html>.
2. The mailing lists used were: arabic-l@byu.edu, corpora@hd.uib.no, and elsnet-arabic@elsnet.org.
3. <http://www.comp.leeds.ac.uk/latifa/>

References

- Abu Shawar, B. & Atwell, E. (2005). A chatbot system as a tool to animate a corpus. *ICAME Journal*, 29, 5–24.
- Ahmad, K. & Corbett, G. (1987). The Melbourne-Surrey corpus. *ICAME Journal*, 11, 39–43.
- Al-Ajmi, H. (2003). Compiling an English-Arabic parallel text corpus. In M. Murata, S. Yamada & Y. Tono (Eds.), *Proceedings of Asian Association for Lexicography* (pp.51–54). Meikai University: Asialex. (<http://www.asialex.org/>)
- Al-Ansary, S. (2003). NP-Structure Types in Spoken and Written Modern Standard Arabic (MSA) Corpora. In D. Parkinson & S. Farwanah (Eds.), *Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics* (pp. 149–180). Salt Lake City.
- 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.
- Al-Sulaiti, L. (2004a). Arabic corpora and corpus analysis tools. In B. Bel & I. Marlien (Eds.), *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles* (volume 2, pp. 547–552). ATALA.
- Al-Sulaiti, L. (2004b). TALN'04: The North African Experience. *ELSnews: Newsletter of the European Network in Human Language Technologies*, 13 (1), 11–12.
- Al-Sulaiti, L. (2003). Computer Assisted Language Learning (CALL): Lille hosts ELSNET's 11th Summer School. *ELSnews: Newsletter of the European Network in Human Language Technologies*, 12 (3), 1–3.
- Al-Sulaiti, L. (1991). *Arabic children's stories*. Ministry of Information and Culture. State of Qatar.
- Al-Sulaiti, L. & Atwell, E. (2004). Designing and developing a corpus of contemporary Arabic. In J. Santana, P. Urena & A. Villegas (Eds.), *Proceedings of TALC 2004: the sixth Teaching and Language Corpora conference* (pp. 92–93). Granada.
- Al-Sulaiti, L. & Atwell, E. (2003). *The Design of a corpus of Contemporary Arabic (CCA)*. School of Computing, Research Report Series, 2003.11. University of Leeds.
- Al-Sulaiti, L. & Knowles, G. (2002). A multimedia Arabic course. In A. Braham (Ed.), *Proceedings of the International Symposium on The Processing of Arabic* (pp. 94–105). University of Manouba, Tunisia.
- Al-Wasiit. (1960). *Al-Wasiit Dictionary*. Cairo: Academy of the Arabic Language.
- ANC. (2003). *American National Corpus*. (<http://americannationalcorpus.org/>)

- An-Nahar. (2004). *An-Nahar Newspaper*. (<http://www.annaharonline.com/>)
- Aston, G. & Burnard, L. (1998). *BNC Handbook: Exploring the British National Corpus with Sara*. Edinburgh: Edinburgh University Press.
- Atwell, E. (1999). *The Language Machine*. London: British Council.
- Atwell, E. (1988). Transforming a Parsed Corpus into a Corpus Parser. In M. Kytö, O. Ihalaainen & M. Risanen (Eds.), *Corpus Linguistics, Hard and Soft: Proceedings of the ICAME 8th International Conference on English Language Research on Computerised Corpora* (pp. 61–70). Amsterdam: Rodopi.
- Atwell, E. (1987). How to detect grammatical errors in a text without parsing it. In B. Mae-gaard (Ed.), *Proceedings of EACL: the Third Conference of European Chapter of the Association for Computational Linguistics* (pp. 38–45). New Jersey: ACL.
- Atwell, E., Al-Sulaiti, L., Al-Osaimi, S. & Abu Shawar, B. (2004). A review of Arabic corpus analysis tools. In B. Bel & I. Marlien (Eds.), *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles* (volume 2, pp. 229–234). ATALA.
- Atwell, E. et al. (2003). The ISLE Corpus: Italian and German Spoken Learners' English. *ICAME Journal*, 27, 5–18.
- Atwell, E., Howarth, P., Souter, C., Baldo, P., Bisiani, R., Pezzotta, D., Bonaventura, P., Menzel, W., Herron, D., Morton, R. & Schmidt, J. (2000a). User-Guided System Development in Interactive Spoken Language Education. *Natural Language Engineering Journal* 6 (3–4), Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, 229–241.
- Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C. & Wilcock, S. (2000b). A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal*, 24, 7–23.
- Bergh, G. (2005). Min(d)ing English language data on the Web: What can Google tell us? *ICAME Journal*, 29, 25–46.
- BNC (2003). *British National Corpus*. (<http://www.natcorp.ox.ac.uk/>)
- Brockett, A., Atwell, E., Taylor, O. & Page, M. (1989). An Arabic text database and glossary system for students. In *Proceedings of the Seminar on Bilingual Computing in Arabic and English* (pp. 154–162). University of Cambridge.
- Brown, J. D. (1988). *Understanding research in second language: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.
- Buckwalter, T. (2003). *Buckwalter Arabic Corpus*. (<http://www.qamus.org/wordlist.htm>)
- Burnard, L. (2004). BNC-Baby and Xaira. In J. Santana, P. Urena & A. Villegas (Eds.), *Proceedings of TALC 2004: the Sixth Teaching and Language Corpora conference* (pp. 84–85). Granada.
- Dury, P. (2004). Building a bilingual diachronic corpus of ecology: The long road to completion. *ICAME Journal*, 28, 5–16.
- Elkhafafi, H. (2001). Teaching listening in the Arabic classroom: a survey of current practice. *Al-^cArabiyya*, 34, 55–90.
- Elliott, D., Atwell, E. & Hartley, A. (2004). Compiling and using a shareable parallel corpus for MT evaluation. In L. Kranias, N. Calzolari, G. Thurmair, Y. Wilks, E. Hovy, G. Magnusdottir, A. Samiotou & K. Choukri (Eds.), *Proceedings of the Workshop on The Amazing Utility of Parallel and Comparable Corpora. Fourth International Conference on Language Resources and Evaluation (LREC)* (pp. 18–21). Lisbon, Portugal.

- Elliot, D. Hartley, A. & Atwell, E. (2003). Rationale for a multilingual corpus for machine translation evaluation. In P. Rayson, A. Wilson & D. Archer (Eds.), *Proceedings of CL 2003: International Conference on Corpus Linguistics* (pp. 191–200). University of Lancaster.
- ELRA (2003). *European Language Resources Association*. (<http://www.elra.info/>)
- Ferguson, C. A. (1971). Problems of teaching languages with Diglossia. In A. S. Dil (Ed.), *Language Structure and Language Use* (pp. 71–86). Stanford: Stanford University Press.
- Francis, W. & Kučera, H. (1979). *Brown Corpus Manual: Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (3rd edition)*. Providence: Brown University. (<http://khnt.hit.uib.no/icame/manuals/brown/index.htm>)
- Ghazali, S. & Braham, A. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. *Arabic NLP Workshop at ACL/EACL*. Toulouse, France. (<http://www.elsnet.org/arabic2001/ghazali.pdf>)
- Graddol, D. (1997). *The Future of English*. London: British Council.
- Green, E. & Peters, P. (1991). The Australian corpus project and Australian English. *ICAME Journal*, 15, 37–53.
- Greenbaum, S. (Ed.) (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.
- Guidère, M. (2002). Toward corpus-based machine translation for Standard Arabic. *Translation Journal*, volume 6, no. 1 (<http://accurapid.com/journal/19mt.htm>)
- Gully, A. (1997). The discourse of Arabic advertising: preliminary investigation. *Journal of Arabic and Islamic studies*, 1, 1–49. (<http://www.uib.no/jais/v001/gully01.pdf>)
- Haddad, S. (1985). Tadris al-mahaaraat al-shafawiyya: mawqif jadiid. *Al-ʿArabiyya* 18, 1–2, 15–21.
- Holes, C. (1990). A Multi-media, topic-based approach to university-level Arabic language teaching. In D. Aguis (Ed.), *Diglossic Tension: teaching Arabic for communication* (pp. 36–41). Leeds: Folia Scholastica.
- Hoogland, J. (1996). The use of OCR software for Arabic in order to create a text corpus of Modern Standard Arabic for lexicographic purposes. In A. Ubaydli (Ed.), *Proceedings of the international conference and exhibition on multi-lingual computing* (pp. 2701–2716). Cambridge University.
- ICE (2003). *International Corpus of English*. (<http://www.ucl.ac.uk/english-usage/ice/>)
- 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 University.
- Izwaini, S. (2003). Building specialised corpora for translation studies. In P. Rayson, A. Wilson & D. Archer (Eds.), *Proceedings of the workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives* (pp. 17–25). Lancaster University.
- Johansson, S., Atwell, E., Garside, R. & Leech, G. (1986). *The Tagged LOB Corpus Users' Manual*. Norwegian Computing Centre for the Humanities, Bergen. (<http://khnt.hit.uib.no/icame/manuals/lobman/INDEX.HTM>)
- Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with Digital Computers*. Department of English, University of Oslo. (<http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM#lob8>)

- Khoja, S. (2003). *APT: Arabic part-of-speech tagger*. PhD Thesis, Computing Department, Lancaster University.
- Leech, G., Garside, R. & Atwell, E. (1983). The automatic grammatical tagging of the LOB Corpus. *ICAME Journal*, 7, 13–33.
- Leech, G. & Smith, N. (2005). Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB. *ICAME Journal* 29, 83–98.
- Lunt, L. G. (1992). Teaching Arabic as a second language in Tunisia. *Al-ʿArabiyya*, 25, 107–125.
- Maamouri, M. & Cieri, C. (2002). Resources for Arabic Natural Language Processing at the linguistic Data Consortium. In A. Braham (Ed.), *Proceedings of the International Symposium on: The Processing of Arabic* (pp. 125–146). University of Manouba, Tunisia.
- McEnery, T. & Wilson, A. (1996). *Corpus linguistics*. Edinburgh University Press, Edinburgh.
- Messaoudi, A., Lamel, L. & Gauvain, J. (2004). Transcription of Arabic Broadcast News. In L. Deng (Ed.), *Proceedings of the International Conference on Speech and Language Processing* (pp. 521–524). Jeju Island. (ftp://tlp.limsi.fr/public/ThA2001o.2_p521.pdf)
- Mili, A. (2003). Teaching (Computer) Sciences in Arabic. In *Arabic Computer Society Newsletter* vol.1. (<http://www.acsportal.org/>)
- Murphy, A. (2005). Markers of attribution in English and Italian opinion articles: A comparative corpus-based study. *ICAME Journal*, 29, 131–150.
- Nicola, M. (1990). Starting Arabic with dialect. In D. Aguis (Ed.), *Diglossic Tension: teaching Arabic for communication* (pp. 42–45). Beaconsfield Papers. Leeds: Folia Scholastica.
- Oostdijk, N. (1988). A corpus for studying linguistic variation. *ICAME Journal*, 12, 3–14.
- Parkinson, D. (2003). Future Variability: A Corpus Study of Arabic Future Particles. In D. Parkinson & S. Farwaneh (Eds.), *Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics* (pp.191–211). Salt Lake City.
- Parkinson, D. (1985). *Constructing the social context of communication: terms of address in Egyptian Arabic*. Berlin/New York/Amsterdam: Mouton de Gruyter.
- Parkinson, D. & Farwaneh, S. (Eds.) (2003). *Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics*, Salt Lake City.
- Peitsara, K. (2004). Variants of contraction: The case of *it's* and *'tis*. *ICAME Journal*, 28, 77–94.
- Peters, P.H. (1987). Towards a corpus of Australian English. *ICAME Journal*, 11, 27–38.
- Rademann, P. (1998). Using online electronic newspapers in modern English-language press corpora: Benefits and pitfalls. *ICAME Journal*, 22, 49–71.
- Sharoff, S. (2004). Towards basic categories for describing properties of texts in a corpus. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, C. Pereira, F. Carvalho, M. Lopes, M. Catarino & S. Barros (Eds.), *LREC04: Proceedings of Language Resources and Evaluation Conference* (volume V, pp. 1743–1746). Lisbon, Portugal.
- Sharoff, S. (Forthcoming). Open-source Corpora: using the net to fish for linguistic data.
- Shastri, S. V. (1988). The Kolhapor corpus of Indian English and work done on its bases so far. *ICAME-Journal*, 12, 15–26.
- Sinclair, J. (1996). Preliminary recommendations on text typology. Eagles Document EAG-TCWG-TTYP/P. (<http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html>)

- Stevens, V. (1993). Concordances as enhancements to language competence. *TESOL Matters*, 2, 6–11.
- Taylor, S. (2003). Comparing Frequencies of Lexical Productions in Arabic Words. In D. Parkinson & S. Farwaneh (Eds.), *Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics* (pp. 181–189). Salt Lake City.
- 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-journal*, 17, 119–124.
- van Mol, M. (2003a). Evolution of MSA, the Case of Some Complementary Particles. In D. Parkinson & S. Farwaneh (Eds.), *Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics* (pp. 135–147). Salt Lake City.
- van Mol, M. (2003b). *Variation in Modern Standard Arabic in radio news broadcasts, a synchronic descriptive investigation into the use of complementary particles*. Peeters. Belgium.
- van Mol, M. (2000a). The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (Eds.), *Proceedings of the ninth EURALEX International Congress* (pp. 831–836). Stuttgart, 8–12 August. (http://www.ilt.kuleuven.ac.be/ilt/arabic/_pdf/stuttgart.pdf)
- van Mol, M. (2000b). *Exploring annotated Arabic corpora: preliminary results*. (http://www.ilt.kuleuven.ac.be/ilt/arabic/_pdf/tunis.pdf)
- van Mol, M. & Paulussen, H. (2001). AraLat: a relational database for the development of bilingual Arabic dictionaries. In S. Lee (Ed.), *Proceedings of Asialex 2001, Asian Bilingualism and the Dictionary* (pp. 206–211). Seoul, August 2001. (http://www.ilt.kuleuven.ac.be/ilt/arabic/_pdf/asialex.pdf)
- Wilson, A. (2005). Modal verbs in written Indian English: A quantitative and comparative analysis of the Kolhapur corpus using correspondence analysis. *ICAME Journal*, 29, 151–170.
- Xu, J., Fraser, A. & Weischedel, R. (2002). Empirical studies in strategies for Arabic retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp.269–274). Tampere, Finland. (<http://www.isi.edu/~fraser/pubs/sigir2002.pdf>)
- Younes, M. (1990). An integrated approach to teaching Arabic as a foreign language. *Al-ʿArabiyya*, 23, 1–2, 105–122.
- Zemanek, P. (2001). Clara (Corpus Linguae Arabicae): An Overview. In ELSNET (Ed.), *Proceedings of ACL/EACL workshop on Arabic language processing*. Toulouse, France. (<http://www.elsnet.org/acl2001-arabic.html>)

Author's address

Eric Atwell and Latifa Al-Sulaiti
School of Computing, University of Leeds
Leeds LS2 9JT, England

E-mail: eric@comp.leeds.ac.uk