The Design of a Corpus of Contemporary Arabic (CCA)

Latifa Al-Sulaiti and Eric Atwell, School of Computing, University of Leeds

Abstract

Corpora are an important resource for both teaching and research. Since Arabic lacks enough resources in this field, 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. A quantitative analysis of the result shows that a number of text types should have priority in the corpus. However, 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, and that there is a need for a corpus of contemporary Arabic covering a broad range of text-types. Our survey also showed support for the inclusion of parallel English-Arabic samples. Supplementary questions showed support for potential use in wide range of Language Engineering applications; and indicated that teachers of Arabic as a foreign language already make significant use of computers in teaching, and want to include contemporary, authentic examples.

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 and American) but other varieties such as Australian (Ahmed & Corbett 1987; Peters 1987), Indian (Shastri 1988), Cameroonian (Tiomajou 1993), and others. 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, 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 a corpus. Nowadays corpora and bilingual parallel corpora in particular, are established tools in MT research and development. This paper discusses the development of a corpus of contemporary Arabic, including parallel English-Arabic samples. 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 MT as well as TAFL materials. Our corpus will be aimed at specific users. Thus designing and compiling the corpus will depend on the views of these 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 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. Table 1 gives a brief description of existing Arabic corpora; these are mostly untagged and restricted to a specific genre.

Table 1. Classification of Arabic Untagged Corpora

Name of Corpus	Type	Source	Medium	Size	Purpose	Material
Nijmegen Corpus (1996)	Academic	Nijmegen University	Written	17 MB	Arabic-Dutch / Dutch-Arabic dictionary	Magazines + fictions
Arabic Newswire Corpus (1994)	Academic	University of Pennsylvania LDC	Written	80,000,000 words	Education and the development of technology	Agence France Presse
Broadcast News Speech (2000)	Academic	University of Pennsylvania LDC	Spoken	More than 110 broadcast		News broadcast from the radio of voice of America.
CALLFRIEND Corpus (1995)	Academic	University of Pennsylvania LDC	Conversational	60 telephone conversations	Development of language identification technology	Egyptian native speakers
CALLHOME Corpus (1997)	Academic	University of Pennsylvania LDC	Conversational	120 telephone conversations	Speech recognition produced from telephone lines	Egyptian native speakers
An-Nahar Corpus (2001)	Commercial	ELRA	Written	140 m words	General research	An-Nahar newspaper (Lebanon)
Al-Hayat Corpus (2002)	Commercial	ELRA	Written	18,639,264 m words	Language Engineering and Information Retrieval	Al-Hayat newspaper
CLARA (1997)	Academic	Charles University, Prague	Written	Up to now 37,000,000 words	Development of machine translation from Arabic to English.	Periodicals + books + internet sources from 1975-present
Classical Arabic Corpus (CAC)	Academic	University of Manchester, UK	Written		General research	www.muhaddith.org and www.alwaraq.com
Egypt	Public	John Hopkins University	Written			a parallel corpus of the Qur'an in English and Arabic

The above corpora are not accessible to the public except the corpus 'Egypt'. In addition to the above list, we know of two other Arabic corpora under progress: one deals with scientific texts (5M words) and the other deals with computer and software texts (2M words) (Personal communication).

The above corpora represent raw material. The only existing Part-of-Speech-tagged corpus consists of 50,000 words based on newspaper texts (Khoja 2002). However, this size of corpus is generally not large enough for research purposes. In order to achieve a reliable result in most 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.

Our survey shows that although there are a number of Arabic corpora available they are mostly built on newspaper and other texts, which use Standard Arabic. In addition, they are not publicly accessible.

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.

For the past twenty years or so there has been a debate over the type of Arabic that is preferable to be taught to foreigners. 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 Elkhafaifi (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'* (ibid).

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.

In a recent article (Mili 2003), the author expressed his disappointment with the level to which Arabic has sunk, especially in the fields of science and technology. Both teaching and research are conducted in English. He rightly fears that Arabic would be approaching the level of extinction if there were no collective effort from the public and private sectors to revive it. 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 will include some technical and scientific material as part of its content, will become one of the resources that contribute to the teaching and exploring of the language of science. In addition, this corpus will be the source we plan to utilise for developing Arabic language teaching materials (Al-Sulaiti & Knowles 2002).

4. Present Project

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 will include texts other than newspaper documents. Our main objective is to make this corpus available for the public, particularly for teaching Arabic as a foreign-language, and for use in both language processing and general 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 categorizations of the topics. However, for this corpus we will make our decision based on the needs and views of end-users. Our target users will be mainly teachers of Arabic as a foreign language and language engineers.

5. Corpus size

Due to limitation of time, our initial target will be to compile a corpus of 1 million words. Our data will be wholly derived from texts received in machine-readable form. Optical scanning of texts will be avoided as it has its limitations and is a very slow process. Nowadays in most Arab countries, publishing companies produce tremendous amounts of material on the computer. Thus there is a growing number of texts available in machine-readable form and we have already identified several promising sources. The sites we have identified contain texts with 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.

Because of shortage of time and funding we are going to take advantage of all these available sources and include as many texts as we can find. Therefore, the size of the texts in each genre will not be limited to a specific number of words as is the case in most corpora. In this regard, our corpus will resemble the design of the American National Corpus (ANC) (Ide 2003).

6. Methodology

Our choice of text types will reflect the needs of the users. In March 2003 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. We developed an online questionnaire and made it available via mailing lists for language teachers and language engineers (arabic-l@byu.edu - corpora@hd.uib.no - elsnet-arabic@elsnet.org). We also sent it to some individual teachers that we know. 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:

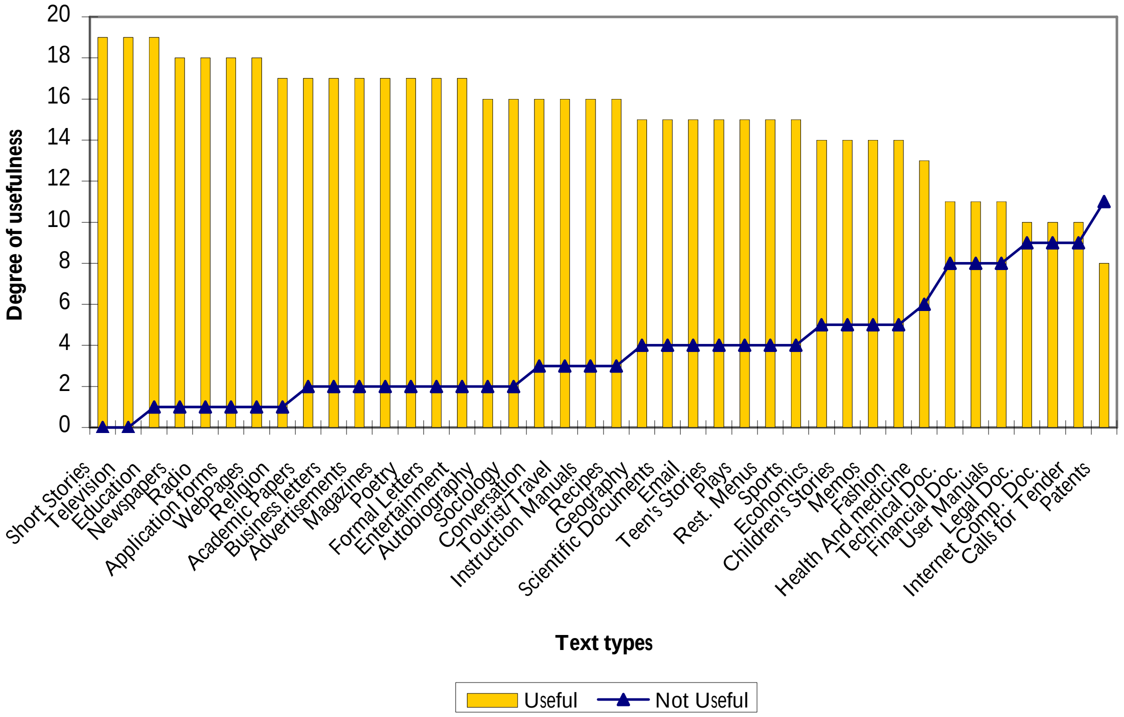
Written: Fiction - Arts - Science - Business- Miscellaneous
Spoken: TV- Radio- Conversation

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

7. Results and discussion

We received 30 replies. (19 from language teachers, 11 from language engineers). We divided the respondents into the two groups and conducted a quantitative analysis using Microsoft Excel. 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. We 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.

Figure 1:Distribution of the useful texts by language teachers



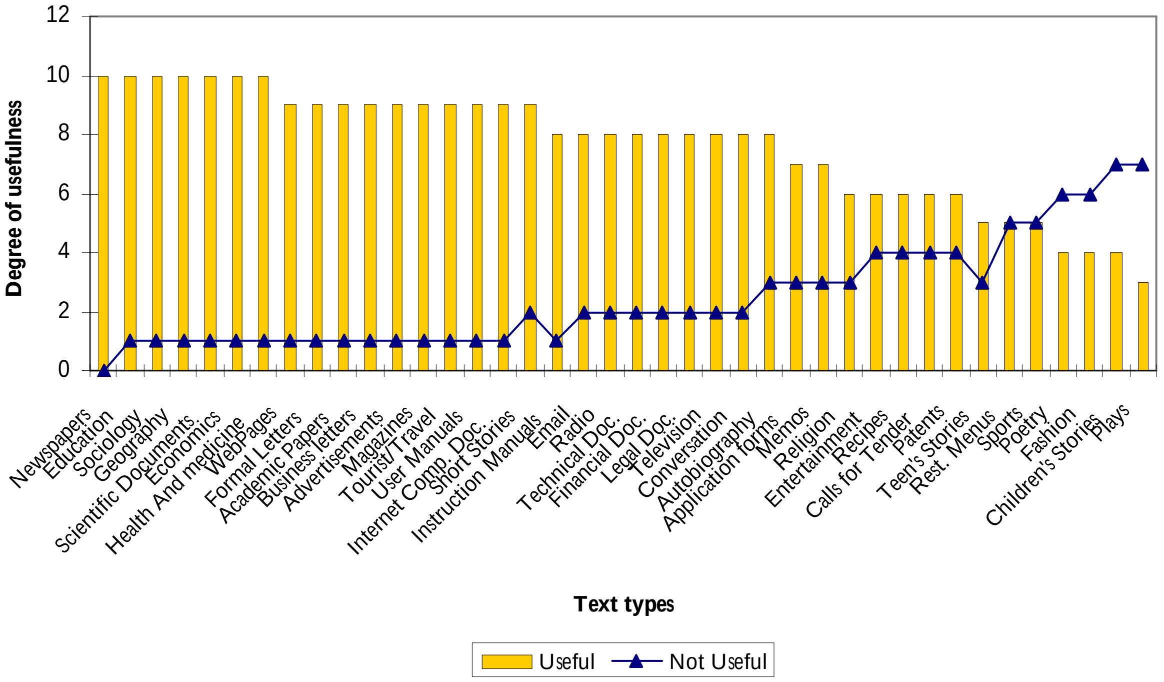
The graph shows that there is an overall consensus over the items ‘short stories’ and ‘television’: none of the language teachers rated these ‘not useful’. The remaining useful texts can be divided into categories based on their usefulness from the point of view of language teachers:

Category 1: 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 doc., 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



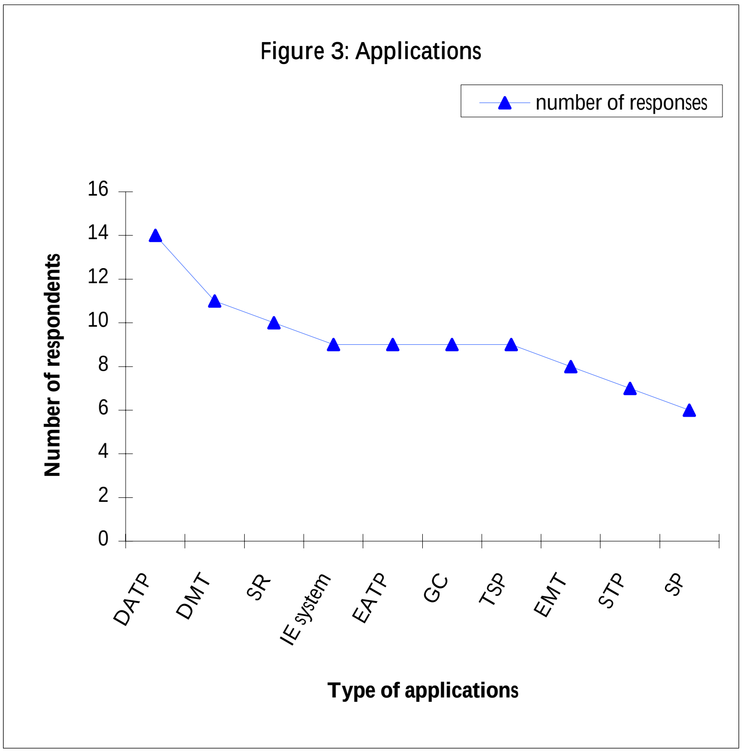
The rest of the texts can be divided into categories according to their classification by language engineers and their value of having equal usefulness. We should point out here that this classification of texts into categories is only made for ease of comparison.

Category 1: education, sociology, geography, scientific doc, economics, health and medicine.
Category 2: web pages, formal letters, academic papers, business letters, advertisements, magazines, tourist/travel, user manual, Internet comp. Doc., short stories.
Category 3: instruction manuals, email, radio, technical doc , financial doc. legal doc, 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 rest. 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 will 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, and that 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 applications we suggested in the questionnaire were:

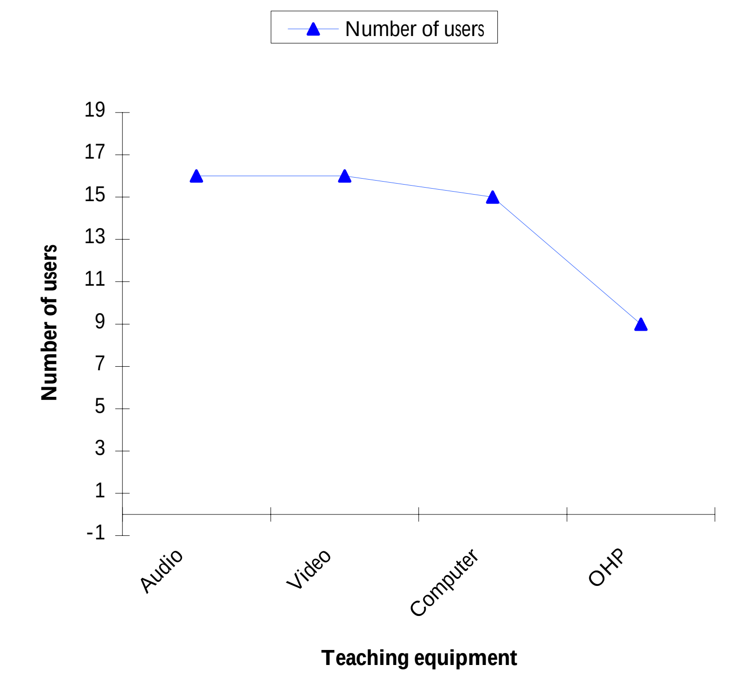
- Developing Machine Translation (DMT)
- Evaluating Machine Translation (EMT)
- Information Extraction systems (IE)
- Developing Arabic text processing systems (DATP)
- Evaluating Arabic text processing systems (EATP)
- Grammar checkers (GCH)
- Speech recognition (SR)
- Speech production (SP)
- Text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

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 table in figure 3 shows 11 respondents out of 15. This is a high score.

This is 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.

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 table 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 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.

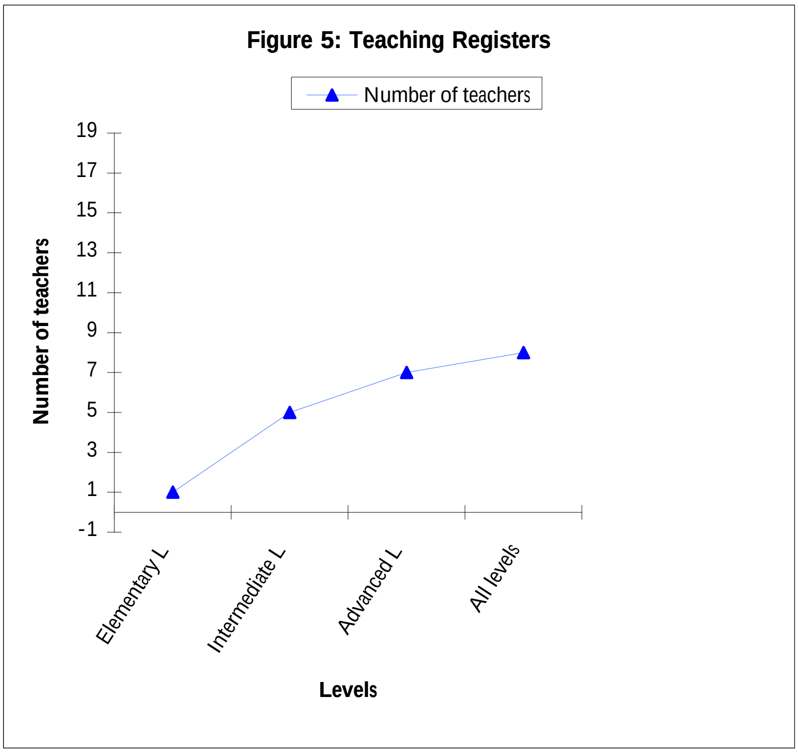


Figure 5: Teaching Registers

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 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’ (1992:122).

8. Limitations and problems of the survey

One limitation of this survey is the small number of replies we received for our research questionnaire. 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. We also encountered some problems when checking the answers. One of these problems was finding some missing answers to some of the questions. We had to obtain the answers by contacting the participants in person.

9. Conclusion

In this report 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 will this corpus fill a gap in the general field of corpus linguistics but it will 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 large data. This corpus should be completed by the end of 2003, but there is a possibility of expanding it at a later date depending on funding.

Acknowledgments

We would like to thank all those who participated in our corpus user survey.

References:

Al-Sulaiti, L. & Knowles, G. (2002). A multimedia Arabic course. In Proceedings of the International Symposium on: The Processing of Arabic, 94-105.

Ahmed, K. & Corbett, G. (1987). The Melbourne-Surry corpus. ICAME 11:39-43.

Atwell, E. (1999). The language machine. London: British Council.

Brown, J. D. (1988). Understanding research in second language: A teacher's guide to statistics and research design. Cambridge: Cambridge University Press.

Elkhafaifi, H. (2001). Teaching listening in the Arabic classroom: a survey of current practice. Al-^{c}Arabiyya 34, pp. 55-90.

Graddol, D. (1997). The future of English. London: British Council.

Haddad, S. (1985). Tadris al-mahaaraat al-shafawiyya: mawqif jadiid. Al-^{c}Arabiyya 18 (1 &2): 15-21.

Holes, C. (1990). 'A Multi-media, topic-based approach to university-level Arabic language teaching' , in Aguis, D (ed.), Diglossic Tension: teaching Arabic for communication, pp. 36-41.Beaconsfield Papers. 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 *Proceedings of the International Conference and Exhibition on Multi-Lingual Computing*, pp.2.7.1-2.7.16.

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).

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.

Lunt, L. G. (1992). Teaching Arabic as a second language in Tunisia. Al-^{c}Arabiyya 25, pp. 107-125.

Maamouri, M & 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*, pp.125-146. Tunisia.

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

Mili, A. (2003). Teaching (Computer) Sciences in Arabic. http://www.arabcomputersociety.org/news/ACSNewsLetterFeb2003Issue.pdf

Nicola, M. (1990). 'Starting Arabic with dialect', in Aguis, D (ed.), Diglossic Tension: teaching Arabic for communication, pp. 42-45.Beaconsfield Papers. Leeds: Folia Scholastica.

Peters, P.H. (1987). Towards a corpus of Australian English, ICAME-journal 11: 27-38.

Shastri, S. V. (1988). The Kolhapor corpus of Indian English and work done on its bases so far. ICAME-journal 12: 15-26.

Tiomajou, D. (1993). Designing a corpus of Cameroonian English. ICAME-journal 17:119-124.

Younes, M. (1990). An integrated approach to teaching Arabic as a foreign language. Al-^{c}Arabiyya 23, pp. 105-22.

Zemanek, P. (2001). Clara (Corpus Linguae Arabicae): An Overview. http://www.elsnet.org/acl2001-arabic.html.