Extending the Corpus of Contemporary Arabic

Latifa Al-Sulaiti and Eric Atwell,
School of Computing
University of Leeds
{latifa, eric}@comp.leeds.ac.uk

Abstract

This paper reports on the development of the Corpus of Contemporary Arabic (CCA), including design, collation and deployment of the initial version, and ongoing work to extend coverage, accessibility, linguistic enrichment, and application.

1 Introduction

Arabic is a major world language, but under-represented in Corpus Linguistics: we surveyed a range of existing corpora, and found them limited in domain, and in coverage of the geographical variation in contemporary Arabic (Hoogland 1996, Zemanek 2001, De Roeck 2002, Maamouri & Cieri 2002, Buckwalter 2003, Izwaini 2003, Elewa 2004). Our impressions were confirmed by a survey of potential users of a corpus of contemporary Arabic: teachers of Arabic as a foreign language (TAFL) and language engineers. This survey guided our design for a new corpus which includes a wide range of sources representing contemporary Arabic. We then set about collecting the corpus, relying mainly on existing sources who granted permission for free reuse of texts. Unlike other Arabic corpora, ours is aimed at TAFL practitioners who are unlikely to have membership of corpus-resource sites such as ELRA or LDC; hence the resulting Initial Version of the CCA has been made available for free use by TAFL teachers and others via WWW: http://www.comp.leeds.ac.uk/latifa

The Initial Version of the CCA is still incomplete, in that we were unable to include all the text categories identified as desirable in the design. In particular, we planned to include a significant proportion of spoken texts, and parallel Arabic-English translated texts; but we were unable to find ready sources, and transcription/translation was too costly and time-consuming for us to attempt ourselves. We are planning a follow-up project to develop an International Corpus of Arabic, following the design of the International Corpus of English.
In this paper we are going to give brief information about the corpus and detail the plan and structure of the follow up Version.

2 Initial version of the corpus

Before collecting the texts, a plan was made to decide on the size of the corpus and text types. In 2003, we conducted a survey of potential users of corpora (e.g.

language teachers and language engineers) to give us guidance for corpus design. The primary aim of the survey was to grade a list of texts on the scale of very useful, useful, and not useful for their own needs. Responses showed that language teachers selected short stories and television to be on top of the list in addition to arts subjects. The science and technical subjects were at the end of the list. However, the language engineers selected newspapers to be on top of the list in addition to a mixture of arts subjects and science.

The collection of the CCA was designed to match the needs of the users. Our initial task was to find online resources and obtain copyright permission from web site owners. Once copyright was granted, we began our text collection. It is worth mentioning that the content of the corpus was more controlled by the availability of the text type on the internet and whether we were granted permission of copyright. This affected the balance of the corpus at this stage in that we found enough material for some text types or even more than enough but little or nothing for others.
Text collection was done manually because only those sites who had given their consent were used. Every text was encoded with a header with the necessary external and internal information and the files were saved as XML documents.

At the end of our 1-year project, we published an initial version of the Corpus of Contemporary Arabic on the World Wide Web[1] . The 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 contains over 843,000 tokens 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 1 shows the text categories which are derived from any of the sources, the number of texts in each category, and number of words. Figure 1 shows the percentage of each source used.

	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

[1] http://www.com.leeds.ac.uk/latifa/

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

Table 1: Number of texts and number of words in each category

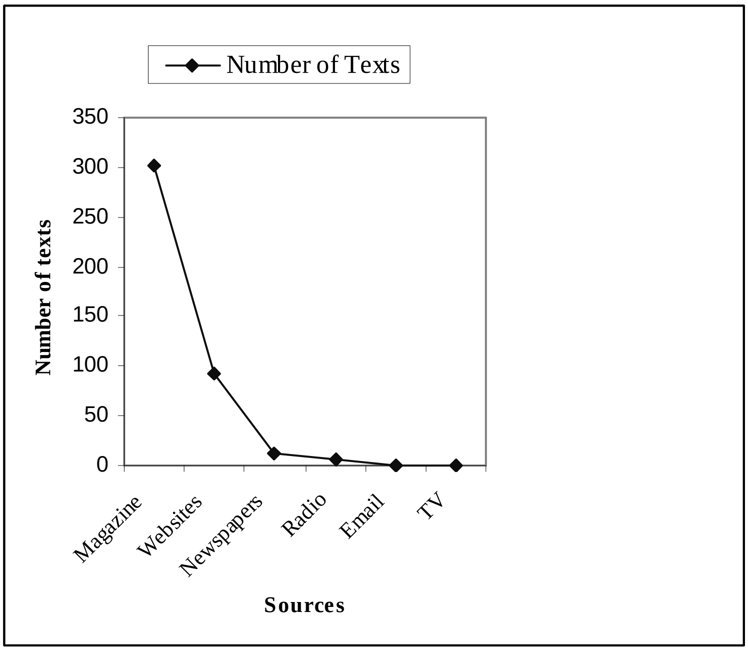

Figure 1: Number of texts collected from the different sources

The original aim was to compile a million-word corpus. However, when the collection phase of the project began, it became obvious to us that this target was unachievable if we consider including a sufficient size of spoken sources and parallel Arabic-English translations. These are not only inaccessible at this stage but they are also time-consuming to process. Instead of filling up the million-word target with printed texts, therefore, we decided to leave a gap for spoken and parallel Arabic-English texts, to be filled later in a follow-up project to extend the corpus of contemporary Arabic to Version 2.

3 **Extending the corpus**

Following the project of the International Corpus of English, we are planning to develop a follow up project to extend the corpus. We will 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. Estimated person-months in the following work-packages refer to Research Fellow (RF) and Software Engineer (SE) time; work by Investigators, Steering Panel members and consultants is not included in these person-month estimates:

3.1 Project Management via ICA International Steering Panel

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. Standards proposals will be drawn up by the project Investigators, but subject to improvement and approval by the International Steering Panel. Membership will include Project Investigators and Researchers Atwell, Al-Sulaiti, and Abu Shawar (Leeds); initial ICA partners in UK university Arabic departments and Arab nations representing the main regional variants, North African Arabic (Jamari and Soudi), Levantine Arabic (Abu Shawar), Gulf Arabic (Abdul-Raof, Holes, Alnajem, and Almuhanna) Saudi Arabic (Ba-Othman, Holes); and representative advisors from Corpus Linguistics and Language Engineering: Khoja (Arabic corpus PoS-tagging), McEnery and Rayson (Corpus Linguistics, BNC and WMATRIX) Burnard (BNC and XAIRA), Nelson (ICE), Choukri (ELRA), Marr (OUP dictionary production and publishing), Beesley (Xerox Arabic language research). The most important function of the International Steering Panel will be to assure the ongoing sustainability of the project both during and beyond its time frame.

3.2 Standards for Corpus Composition and Mark-Up

We propose to adapt the ICE standards for corpus design (Nelson 1996a) and annotation (Nelson 1996b) for the International Corpus of Arabic. This will facilitate comparisons between national varieties of Arabic via national ICA subcorpora, and even comparisons between ICE and ICA subcorpora. The ICA Steering Panel will decide whether and how the adaptation to ICA should allow for some features specific to Arabic, such as the inclusion of religious texts, and English parallel translations. Whereas science and business were major drivers for the internationalisation of English, the international spread of Arabic is more tied up with religion: Arabic is the official language of Islam, and much teaching of Arabic as a foreign language (TAFL) is linked to religion. So, it would seem appropriate to increase the representation of religious sources in the standard design, at the expense of some science and business texts; and to include the Qur'an, the holy book at the heart of Islam, as the first text in the ICA Sampler Corpus. Much official and scientific publication in Arab countries is in English as well as (or instead of) Arabic; for example the first issue of the Arabic Computing Society newsletter (ironically, published in English) entreats Arab computing researchers to publish in Arabic as well as English. So, it would seem appropriate to collect parallel English translations of a proportion of the ICA texts, for use in comparative Arabic-English studies such as Machine Translation evaluation and lexicography.

3.3 Infrastructure for Computer Supported Collaborative Work

We will provide a knowledge management environment for computer-supported collaborative work including discussion and authoring of standards, and download of tools for collation, mark-up, lexico-grammatical analysis, exploration and dissemination of the International Corpus of Arabic. The Virtual Knowledge Park (VKP) environment developed at Leeds University for Computer-Supported Collaborative Work (CSCW) provides a comprehensive range of tools including document management, discussion groups and video conferencing, which enable geographically dispersed team members to share information, discuss project issues and jointly modify team resources, if necessary in real time. The available document management facilities allow for any electronic file to be made available to the wider community in any format. Strict access control mechanisms and version control provides support for managed peer review processes. The VKP can be used as a central storage mechanism for all activities associated with the project. All resources within the VKP (documents, discussion threads etc) can be accessed directly from other web-enabled third party systems through the provision of direct URLs. Links can therefore be built into external systems that will take users directly to resources held in the VKP. Similarly, links to external resources can be embedded within the VKP, thus providing users with a seamless two-way link between information held across different systems.

3.4 Standards for Arabic Morphological Analysis and Part-of-Speech Tagging

We will develop a Part-of-Speech tag-set specification analogous to that of ICE (Greenbaum 1993), but adapted to Arabic morphosyntactic analysis, and conformant to traditional Arab academic models of Arabic grammar. A number of morphological analysers and PoS-taggers already exist for Arabic (see Atwell et al 2004), but generally these are adaptations of PoS-tagging software and/or standards for English. We are planning to develop a new Arabic PoS-tagset combining EAGLES standards for EU languages (Leech and Wilson 1996) with grammar concepts from established academic traditions in Arabic linguistics and TAFL. This constitutes the first draft of the ICA PoS-tagset, to be put to the ICA Steering Panel for refinement and approval.

3.5 PoS-Tagger for the ICA

As already mentioned, a number of morphological analysers and PoS-taggers already exist for Arabic; none of these will be fully conformant to the ICA PoS-tagset specification, but each could provide some features. Our plan is to develop a hybrid tagger merging the analyses of existing taggers, in the style of (van Halteren et al 1998), mapping the output features onto the agreed ICA standard tag-set.

3.6 Collating the ICA Sampler Corpus

We will collect and mark up a 10-million-word sampler corpus as a representative sample of contemporary Arabic. This will involve:

- Searching for and selecting Internet and other freely-available sources from which we can derive our texts, and commissioning the recording and transcription of spoken texts. The latter is labour-intensive, hence time-consuming and costly, requiring a Transcription Fund for fieldwork consultancy; so we will seek to re-use available sources, eg national/local radio/TV broadcasts.
- Selecting and organising the texts, which will include written texts, speech, bilingual Arabic-English texts.
- Obtaining letters of copyright. This will involve in the first place identifying the owners of sources and finding the right addresses.
- Encoding the texts with XML mark-up. Texts with different formats (Doc, PDF, HTML) will be converted into a unified framework (XML format) in which the texts will be enriched with features such as paragraphing and header information regarding text type, author, target audience, etc, following the agreed standards (D1).
- Proofreading the texts; tasks include deleting extra and unnecessary material from texts and checking and adjusting paragraphing markers.

3.7 PoS-tagging the ICA Sampler Corpus

The ICA Sampler Corpus will provide test data for development of the D4 PoS-tagger program. The final version of the program will be used to automatically PoS-tag the full ICA Sampler Corpus; output will also be proofread and manually corrected.

3.8 Tools for Exploring the ICA

As mentioned above, there is a dearth of concordancers or similar corpus exploration tools suitable for Arabic texts. Our research software needs consist of two principal components: text processing tools and text analysis tools. Thus for the former, there will be a need for a lemmatiser, a tagger, and an online dictionary. As for the latter we need a concordance program that generates a frequency list of words, or find relations among selected words, a corpus management program that works as a search engine that searches for a text of specific nature or an author. As far as possible we are going to utilise existing Open-Source software resources with the intention of adapting them to suit the special nature of Arabic. Burnard of Oxford University Computing Service will adapt the XAIRA web-based corpus management and concordancing tool for the British National Corpus (BNC) (Burnard 2004) to explore the Arabic texts of the ICA; this adaptation will then be expanded to internationalization of the XAIRA client interface, to enable delivery of the system via an entirely Arabic interface. Rayson will adapt the WMATRIX corpus-comparison and visualization tool (Rayson 2003) to allow exploration and comparative analysis of ICA subcorpora by linguistics researchers.

3.9 Analysis of lexico-grammatical variation across registers and dialects

In collaboration with International Steering Panel members we will use Arabic concordance and corpus exploration tools to analyse lexical and grammatical variation across the contributing dialects of Arabic in this sampler corpus.

4 Conclusion

This paper gives a brief report on the initial stage of the development of the Corpus of Contemporary Arabic (CCA). We have shown that more spoken texts and parallel texts are important to be included and this could be accomplished in the second stage of development. We have also illustrated the different aspects of the procedure that we would follow. We believe that there is a great demand for such a project for the various benefits that obtained.

English corpus linguistics research has benefited the UK English language teaching industry, including educational institutions and publishers of dictionaries and English language teaching materials; similarly, we predict that the International Corpus of Arabic will give a competitive advantage to UK educational institutions and providers of Arabic language teaching, materials and dictionaries. Longer-term impacts of the work to be done include:

- Cultural: helping to safeguard linguistic and cultural diversity in the information society of tomorrow by strengthening the position of Arabic and other local dialects in linguistics and language technology;
- Social-Political: Promoting cooperation between Britain and the Arab world for the purpose of developing basic components for the multilingual information society;
- Economic: Easing the entrance requirements for British companies into the Arab market and vice versa by providing the basis for automatic translation tools which can be used to translate documentation and marketing material.

Currently we are investigating some European and Arabic funding sources.

References

Al-Sulaiti, L & Atwell, E. (2004) Designing and Developing a Corpus of Contemporary Arabic. In *Proceedings of the sixth TALC conference*. Granada, Spain, p.92.

Atwell, E. Al-Sulaiti, L. Al-Osaimi, S. & Abu-Shawar, B. (2004) A Review of Corpus Analysis Tools. In *Proceedings of JEP-TALN Arabic language processing*, University of Mohamed bin Abdullah, Fez, Morocco, 229-234.

Buckwalter, T. (2003) *Buckwalter Arabic Corpus* homepage. http://www.qamus.org/wordlist.htm. (accessed June 14th, 2005)

Burnard, L. (2004) BNC-Baby and Xaira. In *TALC 2004:* In *Proceedings of the Sixth Teaching and Language Corpora conference*, Granada, 84-85.

De Roeck, A. (2002) ELRA's Al-Hayat Dataset: Text Resources in Arabic Language Engineering. In *ELRA Newsletter* Vol.7 No.1. 2002.

Elewa, A. (2004) Personal communication. UMIST.

Greenbaum, S. (1993) The Tagset for the International Corpus of English, in E. Atwell and C Souter (eds.) *Corpus-based Computational Linguistics***.** (Amsterdam: Rodopi), 11-24.

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, Cambridge,1996,* 2701-2716.

ICE The International Corpus of English http://www.ucl.ac.uk/english-usage/ice/ 2004. (accessed June 12th , 2005)

Izwaini, S.(2003) Building specialised corpora for translation studies. In *Proceedings of the workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, Lancaster University 2003, 17-25.

Khoja, S. (2003) APT: An Automatic Arabic Part-of-Speech Tagger. Ph.D. thesis, Lancaster University.

Leech, G. and Wilson, A. (1996) *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES-Guidelines EAG--TCWG--MAC/R. Final version of 3.1996. EAGLES, Istituto di Linguistica Computazionale, Pisa.

Nelson, G. (1996a) The Design of the Corpus. In Greenbaum, S (ed.) *Comparing English Worldwide: The International Corpus of English* (Oxford: Clarendon Press), 27-35.

Nelson, G. (1996b) Markup Systems. In S. Greenbaum (ed.) *Comparing English Worldwide: The International Corpus of English.* (Oxford: Clarendon Press), 36-53.

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*, Tunisia, 2002, 125-146.

Mili, A. (2003) Teaching (Computer) Sciences in Arabic. In *Arabic Computer Society Newsletter* Vol.1. Available on-line from http://www.tsts.org/mos/content/view/10/2/ (accessed June 13th, 2005)

Rayson, P. (2003) Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University.

Rayson, P., and Garside, R. (1998) The CLAWS Web Tagger. *ICAME Journal* Vol 22, 121-123.

van Halteren et al. (1998) Improving Data Driven Wordclass Tagging by System Combination. In *COLING-ACL'98*, Canada, 491-497.

Zemanek, P. (2001) Clara (Corpus Linguae Arabicae): An Overview. In *Proceedings of ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects.* Available on-line from http://www.elsnet.org/arabic2001/zemanek.pdf (accessed June 13th, 2005)