ABSTRACT for BRISMES Conference 4th – 6th July 2008

Mapping Middle Eastern and North African Diasporas:
Arabic Corpus Linguistics research at the University of Leeds

Eric Atwell, Nora Abbas, Bayan Abu Shawar, Amal Alsaif,
Latifa Al-Sulaiti, Andrew Roberts, Majdi Sawalha

School of Computing, Faculty of Engineering, University of Leeds

http://www.comp.leeds.ac.uk/arabic/

Research on Arabic at the University of Leeds is not confined to the Department of Arabic and Middle Eastern Studies; the Language research group in the School of Computing has an ongoing interest in corpus-based research on Arabic. Language research in Computing is also known as Natural Language Processing, Computational Linguistics, or Language Engineering. Central to our research is the computational modelling of language data; a CORPUS is a text dataset representative of the language to be analysed. We have developed a new Arabic corpus, the Corpus of Contemporary Arabic, freely downloadable from our website (Al-Sulaiti and Atwell 2006, 2005, 2004). Our survey of the needs of teachers of Arabic as a foreign language (TAFL) and Arabic language engineers showed 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 used the World Wide Web to gather a representative selection of contemporary Arabic texts from a wide range of Middle Eastern and North African countries, approximately one million words in total.

We have also developed a new open-source concordance tool for analysis of Arabic corpus texts, aConCorde, also freely downloadable (Roberts et al 2006, 2005); we have explored the use of Arabic concordancer software in Arabic language teaching (Al-Sulaiti et al 2005, 2004). We are also working towards integration of Arabic into the Python Natural Language Tool Kit (NLTK), including software for morphological analysis and Part-of-Speech tagging of classical and contemporary Arabic texts, as represented by the Quran and our Corpus of Contemporary Arabic respectively (Sawalha and Atwell 2008, Atwell 2007, Atwell et al 2004). We have also developed more sophisticated software tools for question-answering and query-by-concept for studying the Quran (Abbas and Atwell 2008, Abu Shawar and Atwell 2005, 2004).

We hope to present our resources to the BRISMES audience, to pinpoint useful applications of these resources in your research, and also to receive requests and/or suggestions for further computational analysis tools which the community might find useful.

BIO: **Eric Atwell** is a Senior Lecturer in the School of Computing at the University of Leeds, where he leads the language research group (http://comp.leeds.ac.uk/nlp). His research interests include language computing, corpus linguistics and machine learning from corpora. He has a BA in computing and linguistics (with ancillary punk rock) from Lancaster University. The other co-authors are current and recent research students in the School of Computing, with research projects related to Arabic Corpus Linguistics:
Nora Abbas: Integrating the Qur'an into the NLTK Natural Language Tool Kit;
Bayan Abu Shawar: A Corpus Based Approach to Generalise a Chatbot System**;**
Amal Alsaif: An automatic analyzer of Discourse Structure for Arabic;
Latifa Al-Sulaiti: Designing and Developing a Corpus of Contemporary Arabic**;**
Andrew Roberts: Grammatical Inference and Corpus Linguistics;
Majdi Sawalha: Part of Speech tagging for Arabic language text.

References (see also http://www.comp.leeds.ac.uk/arabic)

Abbas, Nora; Atwell, Eric. 2008. *A Conceptual Search Tool for the Quran*. Submitted to **COLING'08 International Conference on Computational Linguistics**.

Abu Shawar, Bayan; Atwell, Eric. 2005. *Using corpora in machine-learning chatbot systems*. **International Journal of Corpus Linguistics**, vol. 10, pp. 489-516.

Abu Shawar, Bayan; Atwell, Eric. 2004. *An Arabic chatbot giving answers from the Qur'an* in: Bel, B & Marlien, I (editors) **Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, Volume 2**, pp. 197-202 ATALA.

Al-Sulaiti, Latifa; Atwell, Eric. 2006. *The design of a corpus of contemporary Arabic*. **International Journal of Corpus Linguistics**, vol. 11, pp. 135-171.

Al-Sulaiti, Latifa; Roberts, Andrew; Atwell, Eric. 2005. *The use of corpora and concordance in the teaching of contemporary Arabic* in: **Proceedings of EuroCALL 2005: European conference on Computer Assisted Language Learning**.

Al-Sulaiti, Latifa; Atwell, Eric. 2005. *Extending the corpus of contemporary Arabic* in: **Proceedings of Corpus Linguistics 2005**.

Al-Sulaiti, Latifa; Atwell, Eric. 2004. *Designing and developing a corpus of contemporary Arabic* in: **TALC 2004: Proceedings of the sixth Teaching And Language Corpora conference**, pp. 92-93.

Al-Sulaiti, Latifa. 2004. *The North African Experience*. **ELSNews: Newsletter of the European Language and Speech Research Network**, Vol 13.1, pp.11-12.

Atwell, Eric. 2007. *A cross-language methodology for corpus Part-of-Speech tag-set development* in: **Proceedings of Corpus Linguistics 2007**.

Atwell, Eric; Al-Sulaiti, Latifa; Al-Osaimi, Saleh; Abu Shawar, Bayan. 2004. *A review of Arabic corpus analysis tools* in: Bel, B & Marlien, I (editors) **Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, Volume 2**, pp. 229-234 ATALA.

Roberts, Andrew; Al-Sulaiti, Latifa; Atwell, Eric. 2006. *aConCorde: Towards an open-source, extendable concordancer for Arabic*. **Corpora journal**, vol. 1, pp. 39-57.

Roberts, Andrew; Al-Sulaiti, Latifa; Atwell, Eric. 2005. *aConCorde: towards a proper concordance of Arabic* in: **Proceedings of Corpus Linguistics 2005**.

Sawalha, Majdi; Atwell, Eric. 2008. *Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers.* **Proceedings of COLING'08 International Conference on Computational Linguistics.**