

JEP-TALN 2004 - session on Arabic Language Processing A Review of Arabic Corpus Analysis Tools Un Examen d'Outils pour l'Analyse de Corpus Arabes

Eric Atwell, Latifa Al-Sulaiti, Saleh Al-Osaimi, Bayan Abu Shawar
School of Computing, University of Leeds, Leeds LS2 9JT, England
eric@comp.leeds.ac.uk, latifa@comp.leeds.ac.uk, saleh@comp.leeds.ac.uk,
bshawar@comp.leeds.ac.uk

Résumé – Abstract

Dans ce papier nous présentons une vue d'ensemble critique d'outils disponibles pour l'analyse de corpus arabes, en nous concentrant sur la concordance (Monoconc : Barlow 2003), l'analyse morphologique et le marquage des catégories grammaticales (Shaalán 1989, Ahmed 2000, Khoja 2001, Freeman 2001, Beesley 2001, Berri Zidoum et Atif 2001, Buckwalter 2002, Maamouri et Cieri 2002, Sakhr 2003, Darwish 2003), sur les dictionnaires sous forme exploitable par un ordinateur (Ajeeb 2003) et les outils de visualisation de corpus (Abu Shawar 2004). Nous suggérons qu'une étape essentielle pour les membres de la communauté de recherche d'analyse morphologique arabe devraient se mettre d'accord sur un objectif: choisir un corpus comme échantillon représentatif de textes arabes et se mettre d'accord sur quelles annotations constituent un étalon-or, une sortie correcte décidée que des systèmes devraient viser à répéter. En conséquence, des systèmes individuels peuvent être évalués objectivement en mesurant leurs sorties contre un étalon-or. Le Coran est une norme possible, étant facile à trouver sous la forme de scriptes qui comprennent des signes diacritiques, que des linguistes peuvent enrichir en analyses.

In this paper we present a critical overview of available Arabic corpus analysis tools, focusing on concordancing (Monoconc: Barlow 2003), morphological analysis and Part-of-Speech tagging (Shaalán 1989, Ahmed 2000, Khoja 2001, Freeman 2001, Beesley 2001, Berri Zidoum and Atif 2001, Buckwalter 2002, Maamouri and Cieri 2002, Sakhr 2003, Darwish 2003), Machine-Readable Dictionaries (Ajeeb 2003), and corpus visualization tools (Abu Shawar 2004). We suggest that a vital step for the Arabic morphological analysis research community is to agree a target: select a sample Corpus of Arabic text, and agree on what annotations constitute a "gold standard", an agreed "correct" output for systems to aim to repeat. Then individual systems can be evaluated objectively by measuring how well their output matches the "gold standard" target. The Qur'an is a candidate standard, as it is freely available in vowelised script, for linguists to enrich with analyses.

Mots Clés – Key Words

Corpus, Coran, catégories grammaticales, analyse morphologique, dictionnaire, évaluation; Corpus, Qur'an, Part-of-Speech tagging, morphological analysis, dictionary, evaluation

1 Introduction: Corpus Linguistics and Arabic

Corpora of various types have been developed for a wide range of research and teaching purposes. English, being the main international language, has received the greatest attention among the research community, see eg (McEnery & Wilson 1996, Atwell 1999). Arabic is also an international language, rivalling English in number of mother-tongue speakers. 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 effective corpus analysis tools, specifically concordancers, taggers, morphological analyzers, machine-readable dictionaries, and corpus visualization tools which are necessary for developing and enriching a corpus as a research tool. The Arabic language research group in the School of Computing at Leeds University has surveyed the range of available tools to assist development and analysis of a new Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2004).

The first problem with Arabic which strikes European corpus linguists is the non-Roman script. The Arabic language has 28 consonants and three vowels: a, I, u; these can be short or long, making 6 vowels altogether. Most Arabic words have three consonants, though there can be added prefixes or suffixes. Vowels are “second class letters” in that they are often left out in everyday writing: Arabic texts could be either a vowelised text such as the language of Qur’an or children’s books; or an unvowelled one used in newspapers, books, and media. Handling the unvowelled texts is confusing since the unvowelled word may have more than one meaning. For instance the unvowelled Arabic word “**ك**” (/ktb/, notion of writing) has three possible interpretations: “kataba” (he wrote), “kutiba” (has been written), and “kutubun” (books) (Berri, Zidoum, and Atif 2001).

Until recently, most language processing systems assumed input is encoded in ASCII, which only covers the Roman alphabet. Recent generic software platforms have adopted character-encoding standards which include the Arabic alphabet, including Java’s use of Unicode and Microsoft Word’s utf-8; however, there are still many legacy language-processing systems which have not been extended to include Arabic script.

2 Key-Word-In-Context Concordancers for Arabic text

We investigated a range of concordance programs for Key-Word-In-Context lexical analysis of corpora; the only system which we could find to produce reasonable concordances for Arabic corpus text was Monoconc (Barlow 2003), and even this was not perfect. A common “solution” suggested for some concordancers is a transliteration system: encode the text by replacing each Arabic character with an ASCII equivalent, then extract a KWIC concordance, then re-transliterate the output back to Arabic script. However this constitutes an opaque, user-unfriendly interface: Arabic scholars find it unnatural to treat vowels as full letters in transliteration, and often want to search for roots in terms of consonants only; search patterns must be transliterated too; the output is incorrectly aligned as the right-to-left direction and varying character widths are not handled correctly; and any tags or other annotations added in concordancing are distorted and/or mistransliterated.

3 Morphological analysers and PoS-taggers for Arabic.

For English and French, a first stage in corpus analysis is often Part-of-Speech tagging, also known as PoS-tagging or just Tagging: this involves adding a <tag> to each word indicating its grammatical function in the given sentence, e.g. <simple-past-tense verb> (“wrote”), <perfect-past-tense verb>

(“written”), <plural common noun> (“books”). This generally involves 2 stages: dictionary-lookup and/or morphological analysis to find one or more possible tags possible with each word; and context-sensitive disambiguation, to decide which of the possible tags is best suited to the context, the sentence the word appears in. For English, most words have only one possible tag; and when they have two or more, the program can usually make a choice based on words/tags in immediate context, the tag just before or after the target. For unvowelled Arabic, there are many possible morphological analyses corresponding to alternative vowelings, so tagging is more like full-blown “understanding” of the text. The following is a summary of analysers we found:

3.1 Shaalan’s PROLOG Arabic analyzer

(Shaalan, 1989) reports an MSc project at Cairo University. It is a rule-based system written in SICStus Prolog and needs some background in PROLOG which is difficult to achieve by a typical linguist. It predates modern encoding standards, using an early transliteration scheme.

3.2 Ahmed’s Computational Processor of Arabic Morphology

(Ahmed 2000) is a more recent MSc thesis from Cairo University, reporting on theory and application of a hybrid model of Arabic morphological analysis: “...Morpho3 may be regarded as a demonstration of how a rule-oriented knowledge base and a statistical knowledge base can be married towards solving problems in computational linguistics”.

3.3 Khoja’s APT tagger

APT, Arabic Part-of-Speech Tagger, (Khoja 2001, Khoja et al 2003), uses a combination of statistical and rule-based techniques, as she believes such technique achieves the highest accuracy rates. The tags of the APT tagset are basically derived from the BNC English tagset, modified with some concepts from traditional Arabic grammar. The reason for this is that Arabic has its own unique syntactic, semantic and morphological systems, which make it difficult to adapt to the tagset used for Indo-European languages. The tagset contains 131 tags and they are assigned to words. A corpus of 50,000 words from the Saudi newspaper Al-Jaziira was used to train the tagger.

3.4 Freeman’s Arabic version of the Brill Tagger

(Freeman 2001, 2002) describes another part-of-speech tagger developed for Arabic available on Arabic-1@byu.edu, based on the Brill tagger (Brill 1993), a Machine-Learning system that can be trained with a previously-tagged corpus. His tagset has 146 tags which are assigned to lexemes. This tagset is based on that of the Brown corpus for English. Since this tagset is designed for Indo-European languages, naturally it includes tags for categories that traditional Arabic grammar does not recognize or as belonging to the same categories.

3.5 Beesley’s Xerox Finite-State Morphological Analyser

Beesley (2001, 2003) has developed an Arabic morphological analyser using Xerox generic finite state language-modelling tools. The purpose of this morphological analyser is to use it as a teaching aid and as a component in larger natural –language-processing systems. There is a demo version at <http://www.xrce.xerox.com/research/mltt/arabic>, and a “much improved” commercial version.

3.6 Berri, Zidoum and Atif Morphological analyser

Berri, Zidoum and Atif (2001) developed another morphological analyzer which consists of three main components: a rule knowledge base which has the regular and irregular morphological rules of the Arabic grammar, a set of word lists containing the exceptions handled by the irregular rules, and a matching algorithm that matches the tokens to the rules.

3.7 Buckwalter Arabic Morphological Analyzer

The (Buckwalter 2002) Arabic Morphological Analyzer V1.0. can be freely downloaded from the LDC: <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49> Input text needs to be transliterated to ASCII before processing, and output must be transliterated back to Arabic to be understood. The system does not allow the mixing of Arabic and Roman-alphabet text within the same document; a problem, for example, if text has Roman Part-of-Speech tags or XML markup.

3.8 Maamouri and Cieri LDC Tagger

The Linguistic Data Consortium (LDC) is in the process of developing a POS tagger for Arabic. This tagger is based on the automatic annotation output produced by Tim Buckwalter's morphological analyser of a corpus consisting of 734 files from the 'Agence France Press'. This tagger was developed by Maeda Kazuaki and Hubert Jin.

3.9 Sakhr's Morphological analyser

The Sakhr Company also produced a morphological analyser, which is referred to by Multi-Mode Morphological Processor (MMMP). The Sakhr website <http://www.sakhr.com/> claims that their program covers modern and classical Arabic, and it identifies the base form by removing all the affixes and it gives the morphological pattern. Unfortunately we were unable to get the trial software working, to verify these claims.

3.10 Darwish's Sebawai Morphological Analyser

Sebawai is an Arabic Morphological Analyzer developed by (Darwish 2003) in one day! The morphological analyzer uses Arabic orthographic templates to find roots. "...It's [sic] coverage is not perfect... this morphological analyzer successfully finds the root 84% of the time".

4 Online Arabic Dictionaries

European computational linguists can take for granted availability of high-quality Machine Readable Dictionaries (MRDs) from established sources; for example, Oxford University Press has encouraged corpus-based computational linguistics research by granting Leeds University researchers free access to their online lexical resources including the Oxford-Hachette French Dictionary, the French Source Lexicon, the Oxford-Duden German Dictionary, the New Oxford Thesaurus of English, the Oxford Russian Dictionary, the Oxford Spanish Dictionary, and the Spanish Source Lexicon. However, although OUP also publish Arabic-English dictionaries, these are not in Machine-Readable form: the Arabic text was all hand-written by scribes, so the best we could hope for is scanned images of each page, with no way to convert mixed Arabic and English text images to anything like Unicode or utf-8 text. In fact, most established Arabic dictionaries used by Arabic teachers and scholars are similarly

inaccessible: they predate modern word-processing and electronic publishing methods used currently in European lexicography. At present there are some online Arabic word-translation dictionaries, for example **Ajeeb** (produced by the Sakhr company, see <http://dictionary.ajeel.com/>); but users can only look up one word at a time; it cannot be integrated with Corpus processing infrastructure. There are also other online dictionaries such as **Almisbar**, **Ectaco**, **Lisan Al-Arab**, and **Al-Mawrid** but they are not free to use.

5 Arabic corpus visualization chatbot: learning from the Qur'an

The Quran-28-30 chatbot (Abu Shawar 2004) is software that interacts in a conversation with users using Arabic Language. A java program has been implemented to convert a readable text (corpus) to the AIML chatbot-training format. The conversation generated allows a linguist to “visualize” the corpus in a novel way. The program was tested using the Arabic Qur'an text, which is widely available via the Internet. The result was a version of the chatbot where the user input is Arabic word(s), and the robot response is a list of all ayyas matching this input. It is published using the Pandorobot host service (Pandorobot 2003).

6 Conclusion: the need for a Corpus-based “gold standard”

Our initial aim was to acquire a range of freely-available Arabic morphological analysis systems, to evaluate against samples from our Corpus of Contemporary Arabic under development (Al-Sulaiti and Atwell 2004). It turned out that most systems we could find were either unavailable or very difficult to use, requiring file transliteration or reformatting. Furthermore, there seem to be no agreed standards on what the output of an Arabic morphological analyzer should be: adding vowels to unvowelled text, and/or finding roots and affixes, and/or adding morphosyntactic features, and/or adding Part-of-Speech tags; and if the latter, there is no agreement on appropriate PoS-tagset equivalent to European EAGLES standard.

We suggest that a vital step for the Arabic morphological analysis research community is to agree a target: select a sample Corpus of Arabic text, and agree on what annotations constitute a “gold standard”, an agreed “correct” output for systems to aim to repeat. Then individual systems can be evaluated objectively by measuring how well their output matches the “gold standard” target. This general approach has been adopted in other Natural language Processing comparative evaluations; for example, (Atwell et al 2000) undertook a comparative evaluation of modern English corpus grammatical annotation schemes by applying a range of analysis systems on a standard test corpus, allowing direct quantitative accuracy comparisons. One possibility may be to use the Qur'an as a standard, as it is freely available in vowelled script.

Références

Abu Shawar B, Atwell E (2002). A comparison between ALICE and Elizabeth chatbot systems. School of Computing research report 2002.19, University of Leeds.

Abu Shawar B, Atwell E (2003a). Using dialogue corpora to retrain a chatbot system. Proceedings of the Corpus Linguistics 2003 conference (CL2003), Lancaster University UK. pp. 681-690.

Abu Shawar B, Atwell E (2003b). Machine Learning from dialogue corpora to generate chatbots. Proceedings of BSC-AISIG'03: British Computer Society Artificial Intelligence Special Interest Group, Nottingham

Abu Shawar B (2004), Qur'an chatbot <http://www.pandorabots.com/pandora/talk?botid=94fdd8596e348b29>

Ahmed, M (2000). A Large-Scale Computational Processor of the Arabic Morphology, and Applications. MSc Thesis, Faculty of Engineering, Cairo University, Egypt.

Al-Sulaiti L, Atwell E (2004). The Design of a Corpus of Contemporary Arabic. Submitted to AlArabiyya.

ALICE (2002). A.L.I.C.E AI Foundation , <http://www.alicebot.org/>

Atwell, E (1999). The Language Machine. London: British Council.

Atwell E, Demetriou G, Hughes J, Schiffrin A, Souter C, and Wilcock S (2000). A comparative evaluation of modern English corpus grammatical annotation schemes. ICAME Journal, volume 24, pages 7-23

Barlow M (2003) Monoconc concordancer. <http://www.ruf.rice.edu/~barlow/mono.html>

Beesley, K (2003). Xerox Arabic Morphological Analyzer Surface-Language (Unicode) Documentation. Xerox Research Centre Europe, <http://www.xrce.xerox.com/competencies/content-analysis/arabic-inxight/arabic-surf-lang-unicode.pdf>

Beesley, K (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. Xerox Research Centre Europe, <http://www.xrce.xerox.com/Publications/Attachments/2001-009/finite-state.pdf>

Berri, J, Zidoum, H, & Atif, Y (2001). Web-based Arabic Morphological analyser. In Gelbukh, A (Ed.): CICLing 2001, LINGS 2004, pp. 216-225. Springer-Verlag Berlin Heidelberg.

Brill, E (1993). A corpus-based approach to language learning. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

Buckwalter, T (2002). Arabic morphology analysis. <http://www.qamus.org/morphology.htm>

Darwish, K (2002). Building a Shallow Arabic Morphological Analyzer in One Day. Presented to ACL'02 Workshop on Computer Processing of Semitic Languages, <http://www.cs.um.edu.mt/~mros/WSL/papers/darwish.pdf>

Freeman, A (2001). Brill's POS tagger and a morphology parser for Arabic. In ACL'01 Workshop on Arabic language processing <http://www.elsnet.org/acl2001-arabic.html>.

Freeman, A (2002). What is a word. Proceedings of the International Symposium on the Processing of Arabic, pp. 31-44. Tunisia.

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

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, Pennsylvania.

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

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

Pandorobot (2003). Pandorobot chatbot hosting service <http://www.pandorabots.com/pandora>

Shalan K (1989). Arabic Morphological Analysis and the Lexicon. MSc Thesis, Computer Science Department, Faculty of Computers and Information, Cairo University, Egypt.