Arabic and Arab English in the Arab World

Eric Atwell
Institute for Artificial Intelligence and Biological Systems, School of Computing
Leeds University
eric@comp.leeds.ac.uk

Serge Sharoff
Centre for Translation Studies, School of Modern Languages and Cultures
Leeds University
s.sharoff@leeds.ac.uk

Latifa Al-Sulaiti
Institute for Artificial Intelligence and Biological Systems, School of Computing
Leeds University

Abstract

We begin with two questions about the relative status of Arabic and English in the Arab World: Is there an Arab English? And should Arab science be reported in English or Arabic? To investigate the first question, we collected a WWW corpus of English from Arab countries, and used this as a basis for comparison with UK and US English WWW-corpora. We present the differences found, and possible explanations for the differences. This leads us to some conclusions and ideas for further investigation.

1. Is there an Arab English?

English is widely used as a second language in the Arab world, in education, science, commerce, etc. Computing degree courses in Arab Universities are routinely taught in English – most up-to-date textbooks are in English, imported from USA, UK, and other English-speaking countries. The Arab Open University even directly re-uses English teaching materials provided by the British Open University.

Little formal research has been done on the English used in the Arab world. Is it dominated by British or American English influences? Or is it a recognisable regional variant, on a footing with Indian English or Singapore English?

2. Should Arab science be reported in English or Arabic?

Arab researchers have carried out and reported on their research using either English or Arabic (or both). English is widely accepted as the international language of science and technology, so Arab researchers (along with the rest of the world) must publish in English-language journals and conference proceedings to gain international credibility.

Research papers in Arabic have restricted circulation; for example, ALECSO, the Arab League Educational, Social and Cultural Organisation runs workshops where leading-edge research is reported in Arabic, but this makes the Proceedings inaccessible to the majority English-speaking worldwide research community. In the inaugural issue of the Arab Computing Journal, the editorial urged authors to submit Arabic papers – but this editorial was in English!

Arabic is the first language of most Arabs, but contemporary use shows noticeable variation from Modern Standard Arabic. Some Arab researchers may feel doubly stigmatised, in that their local variant of Arabic differs from MSA, and their English differs from UK or US standard English

3. Collecting a WWW corpus of Arab English

To investigate the English used in the Arab World, we decided to collate a corpus of Arab English. (Al-Sulaiti and Atwell 2006) used WWW sources to gather a representative selection of contemporary Arabic texts. We organised a group of Leeds students to use WWW sources to gather selections of contemporary English texts from a wide range of individual countries. We collected a World Wide English Corpus, analogous to the International Corpus of English used to study national and regional varieties of English. Each student chose one national WWW Top Level Domain, and then used WebBootCat (now part of SketchEngine) to collect approximately 200,000 words of web-page text from a specific country, by restricting the WebBootCat search to the specified national domain.

The resulting World Wide English Corpus is back on the World Wide Web, available for other researchers at http://www.comp.leeds.ac.uk/eric/wwe.shtml
Note that these are "raw" corpus files with no tagging or markup, and some files are formatted idiosyncratically, since not all students followed instructions to the letter ...
From the full corpus of over 20 million words, we extracted a sub-corpus of Arab English: 200K+ word samples from 8 Arab countries: .ae .bh .eg .jo .kw .lb .ma .sa (United Arab Emirates, Bahrain, Egypt, Jordan, Kuwait, Lebanon, Morocco, Saudi Arabia).
The full student collection did include web-text samples from some other Arab countries, but on further investigation we found problems in their format and/or content.

4. Student course-work exercise to write a corpus linguistics research paper

The gathering of the World Wide English Corpus was actually part of a coursework exercise for Computing students. An up-coming approach in Artificial Intelligence and Biological Systems research is agent-based computing: each agent performs a relatively simple task, but many agents combined can achieve complex results. An analogy is a bee-hive: the Queen Bee guides the hive of many simple Workers, and the combined result is a complex, successful system. For this student exercise, the Lecturer (Atwell) was the

"Queen Bee" QB, and the students were the "workers". The exercise followed a pseudo-algorithm, with a complex target outcome: QB+ student co-author a research paper! The following is an outline of the algorithm; The Lecturer role is labelled QB, and the students perform the numbered instructions:

QB) Design the production line: coursework specification:
http://www.comp.leeds.ac.uk/eric/db32/assessment.shtml

QB) Select a domain + research question where Machine Learning is novel:
Language and Cultural studies for a region; specifically:
Which English dominates WWW in this region, British or American?

1) Use AI search tool to choose a region and journal for this question; and find related research to cite, in the Introduction of your paper.

2) Choose 3+ countries in this region, use AI search tool to harvest a Web-Corpus for each country

QB) harvest 10 UK and 10 US Web-corpus data-samples

QB) Use AI tool to find significant differences: candidate ML features characteristic of UK v. US English

3) Choose a small set of features, encode in uk-us ARFF file

4) Chosen region: encode features from (4) in test ARFF file

5) Use AI ML tools (WEKA, log-likelihood etc) to build visualisation and ML evidence of uk-us decision; copy into journal paper: novel evidence (novel for this readership!)

6) Predictions for region samples: UK or US? (Test options: Supplied test set); copy into journal paper

7) Finish paper: Introduction, Methods, Results (ML evidence: novel to this research journal readership), Conclusions

8) Submit paper via AI Knowledge Management tool

QB) assess course-works, aka review/improve

5. Comparison with UK and US WWW corpora of English

At the end of the exercise, we had a collection of student reports, with conclusions about the status of English used in each country. For the 8 Arab countries representing Arab English, the individual student reports found:

English in.eg .jo (Egypt, Jordan) is more like .uk British English

English in .kw .lb .sa (Kuwait, Lebanon, Saudi Arabia) is more like .us American English

English in .bh .ae .ma (Bahrain, United Arab Emirates, Morocco) is like both .us and .uk, showing signs of both British and American English influences.

We then collated all 8 national English samples into a single .ARAB Arab English corpus, and used corpus-comparison software tools to compare .ARAB against UK and US English standards, in a consistent way. Our Log-Likelihood corpus comparison tool produced lists of words markedly more frequent in Arab English. NOTE that a WWW-corpus can only give us lexical differences; differences in accent or pronunciation don't show up on WWW texts.

6. Corpus composition

First, to get an idea of the .ARAB corpus composition, we ran an automatic genre classification program (Sharoff, 2007). This found the following distribution of text genres in the Arab English corpus :

188 discussion
44 information (lists, catalogues, dictionaries)
130 instruction (how-tos, FAQs, tutorials)
99 propaganda (adverts, political pamphlets)
1 recreation (fiction and popular lore),
54 regulation (laws, small print and similar)
52 reporting (newswires, police reports, CVs)

This range of genres is not too dissimilar from the range of genres in other standard English corpora, and at least shows the corpus is not narrowly focussed on a single text type. Fiction and popular lore are noticeably lacking, but this is generally true of other web-sourced corpora; fiction is not so widely found on the WWW. Also literary texts are closely related to one's native language: it is much less likely than an Arab speaker will write fiction in English and publish it on the web.

7. Differences found: UK as reference

We used the Log-Likelihood corpus comparison tool to find words which are relatively more common in Arab English web-text than in British English web-text. The following shows the top findings, the words with most significant difference in frequency:

Word	Frq1	Frq2	LL-score
s	65	7277	9484
Al	17	2063	2699
shall	290	2863	2528
the	85402	96779	2149
Arab	9	1461	1936
Bahrain	3	1400	1905
of	51151	60440	1854
Saudi	11	1252	1633
t	22	1247	1551
Islamic	16	1085	1370

7.1. Differences found: explanations

Most words with high Log-Likelihood scores were names of places, people etc, locally significant – not really a purely LINGUISTIC feature. For example: *Bahrain, Saudi, Islamic.*

Al signifies a possible tokenisation problem in the WWW-trawling program which collected the texts initially; probably many cases of "Al" should be part of a longer word, eg Al-Sulaiti.

Shall is outdated in modern British English, hence rarely used in the .uk sample

the and *of* are slightly overused (misused?)

s and *t* : Arab web-pages tend to use more contractions/enclitics, for example *he's v he is, can't v can not*

We found no clear overall preference for British v American spelling, eg *color* v *colour*, *centre v center*

8. Differences found: Arab as reference

We then used the Log-Likelihood corpus comparison tool to look at the other end of the frequency-differences: to find words which are relatively more common in British English web-text than in Arab English web-text. The following shows the top findings, the words with most significant difference in frequency:

Word	Frq1	Frq2	LL-score
You	4340	9646	1669
BBC	24	1349	1645
UK	136	1542	1333
I	5547	10532	1185
London	119	1176	960
Experience	28	722	786
your	2029	4398	717

8.1. Differences found: explanations

Again, it seems that most words with high Log-Likelihood scores were names of places, people etc, locally significant – not really a purely LINGUISTIC feature. For example, *BBC, UK, London* are more common in British English texts.

you, your and *I* appear to show a preference for interaction with the user in British English web-texts.

9. Possible underlying explanations for linguistic differences

We looked in more detail at a concordance of the Arab English web-pages, and the language used seemed generally less formal, for example:

*This study is not merely used to measure results: it ' **s** a means of giving the employees a voice*

*Perhaps it ' **s** accurate to add that a third, implicit assumption that also influenced the ...*

One possible explanation is that Arabs are more relaxed and informal, like chatting more? Another possible explanation is in the English language education system: perhaps Arab English learners are not taught the way to differentiate between formal and informal usage in their writing in English?

Another common cause of learner English characteristics is L1 influence. Arabic uses clitics much more than English, so it may be that Arabic L1 learners of English carry this tendency over to L2, and naturally use clitics more in their English.

Of course, another possible underlying explanation is that the WWW pages collected may not be representative of Arab English.

10. Conclusions and further work

Natural Language Processing and Corpus Linguistics research has previously focussed on British and American English; is Arab English worth investigating further?

Our collection of World Wide English www-corpus samples is online: http://www.comp.leeds.ac.uk/eric/wwe.shtml

We welcome suggestions applications of these resources in Arab English research, and/or suggestions for extensions to our corpus which might be useful. We want to document the contemporary use of Arab English and Arabic across the Arab world, and develop computational resources for both; and to raise the status of both Arab English and Arabic, so they are recognised as different but equal alongside American English and British English in the Arab world and beyond.

References

Al-Sulaiti, L; Atwell, E. 2006. The design of a corpus of Contemporary Arabic. International Journal of Corpus Linguistics, vol. 11, pp.135-171.

Sharoff, S. 2007. Classifying Web corpora into domain and genre using automatic feature identification. In Proceedings of Web as Corpus Workshop, Louvain-la-Neuve, September 2007.