

# COMPARATIVE EVALUATION OF ARABIC LANGUAGE MORPHOLOGICAL ANALYSERS AND STEMMERS

Majdi Sawalha and Eric Atwell

School of Computing, University of Leeds, Leeds, LS2 9JT, U.K.

## Three Stemming Algorithms

### Shereen Khoja Stemmer

We obtained a Java version of Shereen Khoja's stemmer. Khoja's stemmer removes the longest suffix and the longest prefix. It then matches the remaining word with verbal and noun patterns, to extract the root. The stemmer makes use of several linguistic data files such as a list of all diacritic characters, punctuation characters, definite articles, and 168 stop words.

### Tim Buckwalter Morphological analyzer

Tim Buckwalter developed a morphological analyzer for Arabic. Buckwalter compiled a single lexicon of all prefixes and a corresponding unified lexicon for suffixes instead of compiling numerous lexicons of prefixes and suffix morphemes. He included short vowels and diacritics in the lexicons.

### Tri-literal Root Extraction Algorithm

Al-Shalabi, Kanaan and Al-Serhan developed a root extraction algorithm which does not use any dictionary. It depends on assigning weights for a word's letters multiplied by the letter's position. Consonants were assigned a weight of zero and different weights were assigned to the letters grouped in the word "سائلونيهها". The algorithm selects the letters with the lowest weights as root letters.

## Gold Standard Evaluation Corpus

### 1- Chapter 29 of the Qur'an

### 2-Newspaper text taken from the Corpus of Contemporary Arabic developed at the University of Leeds, UK.

Roots extracted have been checked by Arabic Language scholars who are experts in the Arabic Language.

## Four Accuracy Measurements

### 1- All tokens including Stop words

### 2- Tokens excluding Stop words

### 3-Word Types including Stop words

### 4-Word Types excluding Stop words

## Test Documents

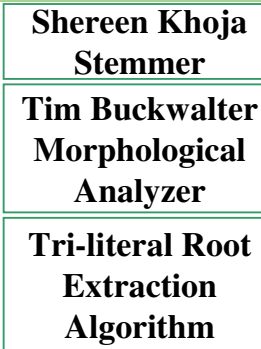
أرجو أن يكونوا قد تعلموا أن الله لا يفتنهم بل يفتنهم من حيث يشاءون. والله ذو العرش العظيم. ومن جاهد نفسه يخافه الله أن الله يعي من العالين وأنهم أتوا ربهما والصلوات تكفلونهم سيئاتهم ويخففون عنهم ما عملوا من سيئاتهم. إنسان ياتيه حسرت وإن جاهدك بشرتك ما تلبس لك به علم ولا نطقها إنى ترجعك لأنك بما كتبت تعلمون. وأنتم أتوا ربهما والصلوات تكفلونهم في الصالحين.

Alif, Lam, Mim. Do men imagine that they will be left (at ease) because they say, We believe, and will not be tested with affliction? Lo! We tested those who were before them. Thus Allah knoweth those who are sincere, and knoweth those who feign. Or do those who do ill-deeds imagine that they can outstrip Us? Evil (for them) is that which they decide. Whose looketh forward to the meeting with Allah (let him know that) Allah's reckoning is surely nigh, and He is the Hearer, the Knower. And whosoever striveth, strive only for himself; for lo! Allah is altogether independent of (His) creatures. And as for those who believe and do good works, We shall requit from them their evil deeds and shall repay them the best that they did. We have enjoined on man kindness to parents; but if they strive to make thee join with Me that of which thou hast no knowledge, then obey them not. Unto Me is your return and I shall tell you what ye used to do. And as for those who believe and do good works, We verily shall make them enter in among the righteous.

Globalization will stay a hot topic of discussion for a long time. In this article, we consider in depth some of the questions raised by new writers who consider globalization as a new lifestyle for the modern man. Taking the lead from America, many writers describe the multi-ethnic and multicultural American life style as the ideal in the new global village where telecommunication, transportation, information systems and the media shorten the distances between disparate groups. Advocates of this point of view look forward to a new modern man, the Cosmopolitan man.

سقى العولة والى وقت بعد معرفة الاسئلة والاصوية ولى هذا الطاق وقت نامل صيغة فى هذه الاسئلة بدأت منذ فترة موجة جديدة من الكتابات تروج للعولمة باعتبارها الشكل الجديد حياة البشر فى ظل العولمة الامريكى معقد من هذه الكتابات يروج للنسب الامريكى معقد الاعراق والثقافات بوصفه النمط الامريكى للحياة فى القرية الكونية الجديدة التى قاربت وسائل الاتصالات والبراصلات ونظم المعلومات ووسائل الاعلام بين اجزائه المختلفة واصبح شرط هذه الطرة بنسب من نوع جديد بشر كوزموبوليتان

## Stemming Algorithms



## Voting

A program to allow "voting" on the analysis of each word; for each word, examine the set of candidate analyses. Where all systems were in agreement, the common analysis is copied; but where contributing systems disagree on the analysis; take the "Majority Vote", the analysis given by most systems.

## Conclusions

Results of the stemming algorithms are compared with the gold standard using four different accuracy measurements.

Khoja stemmer achieves the highest accuracy then the tri-literal root extraction algorithm and finally the Buckwalter morphological analyzer.

The voting algorithm achieves about 62% average accuracy rate for Qur'an text and about 70% average accuracy for newspaper text

The accuracy rates show that the best algorithm failed to achieve accuracy rate of more than 75%. This proves that more research is required.

algorithms are not available freely on the web, and we have been unable so far to acquire them from the authors.

Table 1: Tokens Accuracy of stemming algorithms after testing on Qur'an gold standard

Number of Tokens including Stop words (978 tokens)			
Stemming Algorithm	Errors	Fault Rate	Accuracy
Khoja stemmer	311	31.8%	68.2%
Tim Buckwalter morph. Analyzer	419	42.8%	57.16%
Tri-literal Root algorithm	394	40.3%	59.71%
Voting algorithm	Ex.1: 434 Ex.2: 405	44.4% 41.4%	55.6% 58.6%
Number of Tokens excluding Stop words (554 tokens)			
Khoja stemmer	209	37.73%	62.27%
Tim Buckwalter morph. Analyzer	123	22.2%	77.8%
Tri-literal Root algorithm	279	50.36%	49.64%
Voting algorithm	Ex.1: 266 Ex.2: 229	48.0% 41.3%	52.0% 58.7%

Table 3: Token Accuracy of stemming algorithms. Tested on newspaper gold standard

Number of Tokens including Stop words (1005 tokens)			
Stemming Algorithm	Errors	Fault Rate	Accuracy
Khoja stemmer	231	22.99%	77.01%
Tim Buckwalter morph. Analyzer	596	59.30%	40.70%
Tri-literal Root algorithm	234	23.28%	76.72%
Voting algorithm	Ex.1: 303 Ex.2: 266	30.15% 26.47%	69.85% 73.53%
Number of Tokens excluding Stop words (766 tokens)			
Khoja stemmer	212	27.7%	72.3%
Tim Buckwalter morph. Analyzer	431	60.70%	39.30%
Tri-literal Root algorithm	253	35.63%	64.37%
Voting algorithm	Ex.1: 303 Ex.2: 266	39.56% 34.73%	60.44% 65.27%

Table 2: Word type Accuracy of stemming algorithms after testing on Qur'an gold standard

Number of Word Types including Stop words (616 word types)			
Stemming Algorithm	Errors	Fault Rate	Accuracy
Khoja stemmer	224	36.36%	63.64%
Tim Buckwalter morph. Analyzer	267	43.34%	56.66%
Tri-literal Root algorithm	266	43.18%	56.82%
Voting algorithm	Ex.1: 242 Ex.2: 219	39.3% 35.6%	60.7% 64.4%
Number of Word types excluding Stop words (451 word types)			
Khoja stemmer	155	34.37%	65.63%
Tim Buckwalter morph. Analyzer	251	55.65%	44.34%
Tri-literal Root algorithm	214	47.45%	52.55%
Voting algorithm	Ex.1: 174 Ex.2: 151	38.6% 33.5%	61.4% 66.5%

Table 4: Word type Accuracy of stemming algorithms. Tested on newspaper gold standard

Number of Word Types including Stop words (710 word types)			
Stemming Algorithm	Errors	Fault Rate	Accuracy
Khoja stemmer	232	32.68%	67.32%
Tim Buckwalter morph. Analyzer	431	60.70%	39.30%
Tri-literal Root algorithm	253	35.63%	64.37%
Voting algorithm	Ex.1: 248 Ex.2: 215	34.93% 30.28%	65.07% 69.71%
Number of Word types excluding Stop words (640 word types)			
Khoja stemmer	184	28.75%	71.25%
Tim Buckwalter morph. Analyzer	423	66.09%	33.91%
Tri-literal Root algorithm	224	35.00%	65.00%
Voting algorithm	Ex.1: 252 Ex.2: 195	39.4% 30.5%	60.6% 69.5%