



ISSN: 0067-2904

## Comprehensive Processing for Arabic Texts to Extract Their Roots

AbuSaleh Belal

Ministry of Education, Zarqa, Jordan

### Abstract

Arabic language is a highly inflectional language where a single word can have different forms using a single root with different interpretations. Arabic does not have a standard way to find roots, the reasons for having inflectional language: suffix, prefix and infix Vowels, which built in complex processes. That is why, words require good processing for information retrieval solutions, until now, and there has been no standard approach to attaining the fully proper root. The applications on Arabic words show around 99% are derived from a combination of bilateral, Trilateral and quad lateral roots.

Processing word- stemming levels in order to extract a root is the process of removing all additional affixes. In case the process of matching between a word and Proper names is available, take off the affixes away, according to patterns and rules with reference to root dictionaries.

This research is new series of steps using a new way of affixes' browsing, vowels and Patterns through three stages of stemming. If a match is not found, vowel replacement and patterns readjusted to check, if not, then the word is kept unmodified.

Search engine, indexing, file classification, clustering etc. need developing the root extraction, where the researcher will introduce recommendations and solutions that participate in improving Arabic root extraction. Research applies comprehensive processing on general collection of documents that done gradually to improve the root extraction by 96%.

**Keywords:** Arabic language, root, morphology, patterns, Solid letters (SL), affixes.

## معالجة شاملة للنصوص العربية لاستخلاص الجذور الصحيحة

بلال ابو صالح

وزارة التعليم ، الزرقاء ، الأردن

### الخلاصة

اللغة العربية هي لغة كثيرة الأوجه حيث يمكن لكلمة واحدة أن يكون لها أشكال مختلفة من نفس الجذر ولمعاني مختلفة. من غير المتوقع أن تحتوي الكلمات العربية على طريقة فريدة لاستخلاص الجذر، والسبب الكلمة يضاف لها: اللاحقة، والبادئة، والداخلية والتي بنيت في عمليات معقدة. ولهذا السبب، تتطلب الكلمات معالجة جيدة لحلول استرجاع المعلومات. حتى الآن، لم يكن هناك نهج قياسي للوصول إلى الجذر الصحيح بالكامل. تظهر التطبيقات على الكلمات العربية أن 99 ٪ مشتقة هي مزيج من الجذور الثنائية والثلاثية والرباعية.

معالجة مستويات الكلمة من أجل استخراج الجذر هي عملية إزالة كل الزوائد الإضافية. في حالة توفر عملية المطابقة بين الكلمة والأسماء الصحيحة، يتم تجريد الزوائد، وفقاً للأنماط والقواعد وايضا مطابقة إلى قاموس الجذور.

تعتمد هذه السلسلة خطوات وبعض التغييرات باستخدام طريقة جديدة لعرض الزوائد وحروف العلة والنمط من خلال ثلاثة مراحل. إذا لم يتم العثور على تطابق، استبدل حرف العلة، وإعادة ضبط النمط، غير ذلك، عندها تبقى الكلمة كما هي. سيتم تشغيل جميع هذه الخطوات من خلال خوارزمية تطبيق في برنامج حاسوب. محرك البحث والفهرسة وتصنيف الملفات والعنقود.. الخ تحتاج إلى تطور أكثر لاستخراج الجذور، وسيقوم الباحث بتقديم التوصيات والحلول التي تشارك في تحسين استخلاص الجذر. البحث يقدم معالجة شاملة لمجموعة مستندات محكمة متنوعة وتم بشكل مدروس لتحسين استخراج الجذر بنسبة 96 %.

## 1. Introduction

Searching the world of technology requires many efforts towards Information Retrieval success. For Text Mining TM on Arabic, some works have been conducted to investigate the effect of using stems or roots instead of words on Text Classification (TC) performance. By comparing the Arabic language with other languages, we find weak in the Arabic text technologically that we seek to develop and improve. This research work to develop the technologies of Arabic root extraction. The automatic extract of the root of the Arabic texts is important match with other techniques:

1. Most researches have to work with Arabic text for trilateral [1] and quad lateral [2] but not with all levels of bilateral, trilateral, quad lateral, pent lateral, six lateral roots.
2. Root extraction gives a valuable support to many natural language processing applications such as information retrieval. Stemmer eliminates the longest suffix and the longest prefix. It then matches [3] the remaining word with verbal and noun patterns.
3. Since such corpora provide researches with resources for performing different computational tasks in many fields of informatics linguistics, such as roots dictionary, information retrieval [4], and more.
4. Document's Categorization: using the Arabic root extraction algorithm to obtain the relevant root helps in classification.
5. Studying this singularity in terms of estimating associative relationships between roots, and all their potential occurrences with multiple morpho-phonetic patterns [5].
6. Improving the performance of search engines by the entire corpus, and a post-retrieval document browsing technique. The amount of documents managed to represent the accumulated knowledge of the organizations is also quickly growing and an efficient access for these documents to become vital. These require a new development in technology fields like:
  - a. Searching: querying and Ranking of search results.
  - b. Optimizing information representation and storage [6].
  - c. Classification and Clustering of Documents [7]

Search Engines, (Google, Yahoo ...) are the common gateways to the huge collections of electronic text and relevant information. In comparing other languages with Arabic in relation to information retrieval, Arabic needs more and more efforts to match with the developed languages.

## 2. RELATED WORK AND BACKGROUND

This research works on achieving some improvement in order to extract the Arabic root, and the processes of extracting the root can be done through various steps. Nevertheless, Arabic language is not ruled affixes that makes its morphological analysis processes hard.

The processes must be organized to the three types of affixes. Prefixes are the extra letters added to the beginning of the word, infixes are the extra letters added to the middle of the word and suffixes are the extra letters added at the end of the word.

### Background

Arabic language is described as algebraic language that makes its morphological analysis process very difficult [1, 8, 9]. It is considered as a very challenging language due to: its complex linguistic structure in which it is characterized by a complex situation and its highly derivational nature where morphology plays a very important role.

Arabic words are divided into two types. Noun and verb are derived from a closed set of words which are generated from a set of roots, which is used now with modern language around 7,000 roots and with traditional language that does (past) not exceed 10,000 roots. Arabic language can build 168,924 roots.

The levels of roots depend on number of letters for each word it distributed approximately as follows:

8 %: Two letter roots (bilateral).

69%: Three letter roots (trilateral).

21%: Four letter roots (quad lateral).

Less than 2%: Five and six letter roots (Pentalateral).

The problem of the study launched in the importance of scientific technologies, sought by the radical solutions to give help in the development of Arabic language technically. Applying the morphology inflection only is not to deal with all affixes and not classified as appendages scheduling, means that there are tens of words which will not find a way the root within the selected text. On the other hand, the diacritics marks play a big role to distinguish the word meaning and the extraction may give different root.

#### **Related Works:**

Considerable research on root-base, stemming stages and morphological analysis [1] is amassing for the Arabic language, no standard root extraction oriented algorithm has yet emerged.

Mohammed Aljlayl [8] process many variant word senses are based on an identical root; thus, the root-based algorithm creates invalid conflation classes that result in an ambiguous query which degrades the performance by adding extraneous terms. This automatic rule-based stemming algorithm is not as aggressive as the root extraction algorithm.

Larkey [1] suggest four different approaches to Arabic stemming, which can be identified – manually constructed dictionaries, algorithmic light stemming which remove prefixes and suffixes, morphological analyses that attempt to find roots, and statistical stemmers, which group word variants. Khoja[9] has developed an algorithm that removes prefixes and suffixes, all the time checking that it's not removing part of the root and then matches the remaining word against the patterns of the same length to extract the root, but did not take into account most of Arabic morphology.

Al-Fedaghi and Al-Anzi [10] algorithm tries to find the root of the word by matching the word with all possible patterns with all possible affixes attached to it.

Anne n.de roeck and waleed al-fares [11] produce a research about a morphologically sensitive clustering algorithm for identifying Arabic root, that depend on clustering the Arabic word into various category and then it will be easy for us to rank it find the appropriate root.

Beesley [12] algorithm removes the longest possible prefix, and then extracts the root by checking the first five letters of the word. This algorithm is based on an assumption that the root must appear in the first five letters of the word.

Abderahim [13] present system composed of two modules, first one consists of an analysis out of context to segment each word of the sentence into its elementary morphological units to get root. The second use the context to identify the correct root among all the possible roots of the word, which give good results.

Dilekh [14] propose to remove the prefixes before suffixes to reach the stem, apply the patterns list and without patterns matching. This method not always give better result.

### **3. Text Collection and Function Root List**

#### **Document Collection**

In order to support and test the work 350 Arabic texts are collected ( about 33,800 word ) arbitrarily chosen from various online Arabic newspapers, magazines, academic and other sources published online, according to ten subject categories: Politics, Economics, Human Rights, ,Sports, Science, Health, arts, Agriculture, Law, and Religions texts. There is no determination of words or text, the algorithm apply generally to any document or collection.

#### **Root lists**

The lists of the root that built by more than 99% of the root type from two to six letters. The root list divided into all parts of root as follows:

Two letter roots (bilateral). "Shada marks(ّ)" mean repeat the letter.

Three letter roots (trilateral), around 3/4 of the roots.

Four letter roots (quad lateral).

Five letter roots (pent lateral).

Six letter roots. (six lateral)

These all root classifications that give more appreciate to extract affixes and match exact root from the lists, which help to improve the percentage of extract proper root.

### **4. Root Extractor Approach**

Arabic contains three genders (much like English), masculine, feminine and neuter. It also contains three persons, one to describe the speaker, one to describe the person being addressed and one to describe the person that is not present. Arabic [9] differs from Indo-European languages in that it contains three numbers instead of the more common two numbers. So as well as singular and plural, there is also the dual that is used for describing the actions of two people. We have based our Arabic tag set on this system of language teaching. As such we distinguish between the three moods of the verb (indicative, subjunctive, and jussive).

### Solid Letters (SL)

It represent the essential letter of the word and not to remove it at all from the word, these letters if they appear as first letter that mean there is no prefixes in word and they appear as last letters that mean there is no suffixes in this word.

This distinguishes between letters give direct extraction to determine the exact root without any doubt of the root. Our algorithm begin by removing light prefix (LP) then all suffixes then remain prefixes, where each step trying to match with suggestion solid letter SL.

The solid letters (SL) are "16" and it is not affixes at all, which mean these letters are essential of the root. Shown as follow

**Table 1-Solid Letters (SL)**

ز	ر	ذ	د	خ	ح	ج	ث
هـ	ق	غ	ط	ظ	ض	ص	ش

Extra condition: Next letters be Solid if follow by any prefix like *وال، بال، عال، ب، كال، ك، ال* these letters are:

**Table 2-Solid Letter after light prefix**

و	س	ن	ل	ع	ب	ك	ف
---	---	---	---	---	---	---	---

Suffix: All letters are solid in root and no suffix if not the end letter is:

**Table 3**

ي	ك	ت	ه	ا	ن	م
---	---	---	---	---	---	---

Infix: Only the infixes letters are Vowels beside "ت":

**Table 4- all suffix solid except table**

ي	و	ت	ا
---	---	---	---

The algorithm of solid letter (SL) is:

If the word does not start by letter of Table-1 and condition of Table-(2) then there is no prefix.

If the word does not have middle letters of Table-(4, 5) then there is no infix.

If the word has not end by letter of Table-3 then there is no suffix.

### Light Prefix (LP)

Light prefixes LP it is part of prefixes where it play role in removing general prefix always be extra letter in root. Removing these LP at the beginning of algorithm give better result than starting by remove all prefixes or suffixes. The Light Prefixes LP is:

**Table 5-Infix letters**

ال	وال	بال	كال	فال
----	-----	-----	-----	-----

The LP needs rules to exclude it, if the word greater than six letters directly remove suggest ( *و، ل،* ) ( *ك، ب، ا، ت، م، ن، ي* ). The previous rule depends on letters numbers from one to four prefix also.

### The Approach

Until now, there is no standard algorithm to extract all the roots from the text. This potential is to enhance the algorithm root extraction and to make it more accurate. The research was worked on the

rule-based root extraction approach and morphological analysis to extract the root of a given Arabic word, stem levels, and most popular patterns.

The powerful of this research is applying general documents (collection). Developing the algorithms that has been proposed previously is to develop and make some experiments on them it also improves some of its parts by constructing a computer program that simulates the Arabic texts. Reformulation of the letters arrangement into Solid, Prefixes, Infixes and Suffixes letters, processing any word depends on the Place of Affix. The affixes remove or/and patterns apply will executing as: (each step compare with root dictionary)

- 1- Remove from text [1, 15] non-Arabic letters and one letter.
- 2- There are three levels of stemming which work on removing affixes to get root:
  - a) Stem1: Removing light prefixes LP.
  - b) Stem2: Removing all suffixes, remain prefixes.
  - c) Stem3: Removing infixes.
- 3- Comparing the patterns [5, 9] with the word.

### 5. Experimental Results:

Root extraction in general collection is the orientation to build the algorithm. Arabic language is a highly inflectional language where the predefined morphological patterns attempt to find root.

Root is necessary in IR; it allows combining terms that have similar meaning, with small differences in the morphological form having a single root, and therefore it improves the quality of retrieval.

The algorithm shows the effectiveness to deal with affix by place. The procedures start by removing most prefixes with applying SL, and then operate suffixes before remaining prefixes during the stemming processes. Stage three of stem to extract the root by removing the infixes and may use replacement between vowels.

Many root algorithms have been developed for a wide range of root extraction. The aim of our experiments is to evaluate different methods of proper root. A series of experiments was conducted to show the effect of each stage of stemming step, morphology inflection, replacement of infixes vowels and Solid Letter finally to get proper root from the word..

Solid letters (SL) help to find the root directly. Remove non-SL in return to dictionary of root and how to exclude the letter of root. As examples, the word "تراحم" the letters ( ر ، ح ، م ) are solid and give the root (رحم), "عنوان" the letters " ن ، ن ، ع " are solid and give the root (عنن), "يندارسون" the letters " د ، د ، ا " are solid and give the root (درس), "العملاقة" the letters " ق ، ل ، م ، ع " are solid and give the root (عملق). The SL algorithm gives direct solution to extract proper root by: determine the solid letters from Tables-(1, 2) and it depend letter place of Tables-(3, 4).

Replacement between vowels ( ا ، و ، ي ) play role to find the root where vowel may be essential of root. The word if not root and have vowel, then algorithm apply replacement to get proper root. The letter priority replacement [16] to compare the root dictionary is letter "و" to try replace with "ا، ي" in any place of suggested root. If not, then priority of replacement first and last letter of root to "ا" to try replace with "ي، و" and compare with root dictionary, second letter, it is better for "ي" to try replace with "ا، و". Also, Replace "او، يو" or "وا، وي" by "و" if it is not first letters in word or the word greater than four letters.

There are many of words when algorithm does not give the proper root, more than one root or wrong root. The problem appears of the same word gives different roots and vice versa. As ex, "بعد" the root may "عود" mean back and music machine, "عدد", mean number and repair, "وعد" mean promise.

Shada(ّ) upper letter, mean extra letter" repeat the letter", but most texts do not use Shada specially of electronic text. The bilateral word needs to add letter to be root. Hundreds of roots need to process, back word to bilateral then add extra letter. Ex, "يحبهم" back to "حبّ" then add extra second letter become "حبيب", "الامتداد" back to "مدّ" then "مدد".

The algorithms process nouns with suggestion table lookup. The contents of the table are names, cities, places, things and so on. Generally, the noun affixes are at the beginning (prefixes) and the end (suffixes) of nouns. The algorithm can only remove LP ( كال ، وال ، فال ، ) and suffixes like ( ي، ية ، ة ) to exclude the noun from the list.

All previous suggestion can compare between words of different other algorithms khoja[9], Ghwanmeh[2] are mentioned in Table-6.

**Table 6-**Comparisons with other algorithms

Word	Khoja	Ghwanmeh	Our Alg.	Comment
بخلاء	خلا	خلل	بخل	S.L.
الجزائرية	--	--	جزائر	Noun
وجودنا	جود	جود	وجد، جود	S.L.+ Rep.
والعماني	--	--	عمان	Noun
السجود	جود	جود	سجد	S.L.+ L.P.
بادرو	درا	بادرو	بدر	S.L.
ورنیش	رنش	ورنیش	ورنش	Quad lateral
يرتقبن	رتق	رقق	رقب	S.L
بحب	بحب	حبيب	حبيب	Bilateral
يسمسرون	مسر	مسر	سمسر	Quad lateral
الفرزدق	فرز	--	فرزدق	pent lateral

With all previous suggestions, the word sometimes may not refer to the proper root. Many of these reasons play an active part to cheat extraction root. The reasons are:

1. The massive number of rules.
2. Diacritics of one word give different roots.
3. Affixes complexity.
4. The conflict of anomaly rules.
5. One root may represent different meaning

Next mention Table-7 show the types of experiment documents and give proper root. The Figure-1 and Figure-2 show the rate of root extraction and proper root.

**Table 7-** types of documents and proper roots

Field	No. of Doc.	No. of Word	Root	Proper root	Proper root Rate
Politics	30	3020	2050	1987	97%
Economics	20	1580	1150	1115	97%
Human Rights	10	1140	520	493	95%
Sports	30	2900	1910	1790	96%
Science	25	2200	1130	1040	92%
Health	25	2870	1520	1444	95%
Arts	15	1730	1150	1115	97%
Agriculture	20	2200	1340	1299	97%
Law	10	1100	950	890	94%
Religions	25	2360	1760	1691	96%
Total	225	22100	13480	12864	96%

Our algorithm used to extract proper root by removing non-root word, making normalization by replacing some letters built for vowels. Determine table of Solid Letters (SL) from tables 1 to 4.

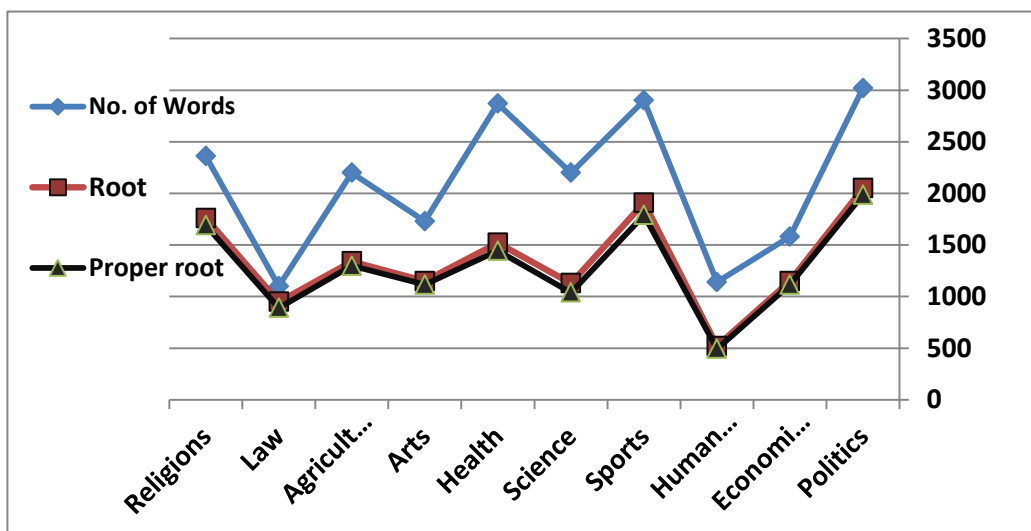


Figure 1-Improve of proper roots

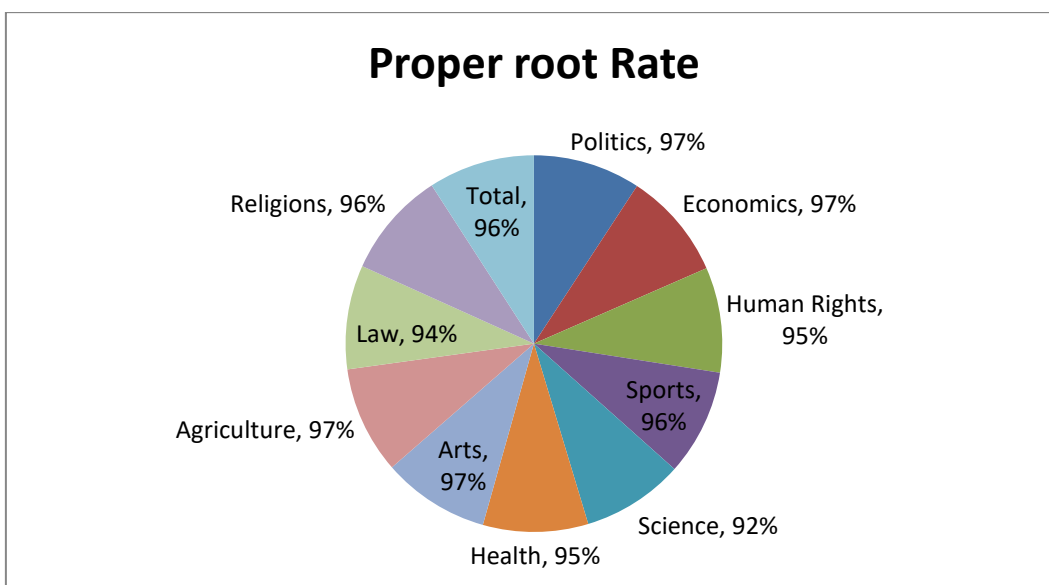


Figure 2-rate of proper roots in each category

The affixes draw in appendix. Generally build algorithm gradually up to remove extra letters from root, as show the algorithm:

1. Remove punctuation, diacritics, numbers, stop words, not Arabic words, and special characters.
2. Normalization:
  - a. Replace the letter (ئ، ع، ه، ا، آ، و، ث) with (ا).
  - b. Replace the duplicate letters by one letter if greater than 3.
  - c. Replace (وا، او، وي، يو) with (و).
  - d. Remove one letter, two letter if not match bilateral root.
3. Determine the SL, Table-(1 to 4), exit if done.
4. Remove prefixes and consider LP (“ال، وال، بال، كال، فال، الا، والا، لل، للا، ولل”). Refer to long of word.
5. Match with noun table, remove suggested prefixes and suffixes.
6. Each step match root:
  - a. remove light prefix LP and apply SL.
  - b. remove all suffixes then apply SL, remain prefixes then apply SL.
  - c. remove infixes (ا، ت، و، ي) then try replacement.

- d. apply replacement vowels
7. Else, return the original word.
8. end

With all previous suggestions, the word sometimes may not refer to the proper root. Many of these reasons play an active part to cheat extraction exact root refer to one or more:

1. The huge number of rules may needed to find root.
2. Diacritics of one word give different roots.
3. Affixes complexity.
4. The conflict of anomaly rules.
5. One root may represent different meaning.

### Conclusion

Many Root algorithms can be employed in Arabic Text Pre-processing to give root. The complexity is the nature of Arabic language. We can enhance our algorithm of exact root by:

Apply all level of root (two to six letters).

Change methods of excluding affixes and replacement between vowels.

Deal with solid letter SL for direct extraction.

Remove LP, Suffixes, Prefixes then Infixes "put in order" give better result.

The Tables-(1, 2, 3, 4, 5) indicators for word have uniform affixes.

Using the pattern if not extract the root by table lookup.

Expert judgment it does used to evaluate the results. The algorithm extracts successfully the proper root with an accuracy rate up to 96%. The algorithm power is using new root extraction methods.

The research proves that it is difficult to extract fully of appropriate root for general collection. The reasons return to the huge number of rules, diacritics of one-word give different roots, affixes complexity and the conflict of the anomaly rules.

Future work needs more help from specialist experts' of Arabic language to improve percentage and accuracy of the appropriate root from word.

### Corpus size and Computer program

The Arabic collection 350 documents are collected (about 33,800 word) arbitrarily chosen from various online Arabic newspapers, Our proposed stemming algorithm using Visual Basic programming language. The system accepts a text file that includes the Arabic words and produces the roots of those words. Examples of the systems' output results shown in Table-(6, 7).

### Appendixes:

#### Table-Prefixes

Start by	The letters
Meem	م، مب، من، مس، مست
Seen	سن، ساء، سي، سييت، ست، سيس، ستس، ستس، ستت، سييت، سنست
Ya'	ي، يت، ين، يست
A	ا، اا، الان، ان، اي، ات، است، الت، الم
Lam	ل، لي، لت، لن، لي، لا
Kaf	ك، با، كا، كالا
ba'	ب، بال، بالا، باست
noon	ن، نت، نست
ta'	ت، تست، تن، تت
fa'	ف، فال، فاء، فلي، فلت، فس، فست، فسي، فسن
+ (و، ال، لل، كال، بال، فال، الا) All letters	

#### Table-Infixes

Vowels	ا، و، ي & ت
--------	-------------



**Table-Suffixes**

One letter	ن، ا، ت، ة، ه، ي
Two Letters	وه، ين، اة، تم، ته، نن، ني، هن، ون، يه، ان، تك، تة، كم، نك، نه، هم، وك، اه، تن، تي، نا، نة، ها، وا، وة، يك، يا، ية
Three Letters	تين، كههم، نيه، نههم، ونه، وها، بههم، ونا، ونك، وني، وهم، تكم، تنا، تها، تني، تهم، كما، كها، ناه، نكم، هما، تان
Four Letters	كموه، ناها، ونني، ونهم، نكما، تموه، تماه، كماه، ناكم، ناهم، نيهما، وننا

**Acknowledgement**

The author would like to thank all persons for their help and insightful suggestions during the conduct of this research.

**References**

1. Leah, S., Larkey, L. B. and Margaret, E. C. **2002** "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis".
2. Sameh, G., Saif, R., Riyad, Al. and Ghassan, K. **2009** "Enhanced Algorithm for Extracting the Root of Arabic Words," Sixth International Conference on Computer Graphics, Imaging and Visualization, IEEE.
3. Brahimi, B., Touahria, M. and Tari, A. **2016**. Data and Text Mining Techniques for Classifying Arabic Tweet Polarity. *Journal of Digital Information Management*. **14**(1): 15-25. 2016.
4. Baranyi, P., Csapo, A., Sallai, G. **2015**. Cognitive Infocommunications (CogInfoCom); Springer International Publishing: Basel, Switzerland.
5. Bassam, H., Ahmad A., Mamoun, H. and Ammar H. **2018**. "Associative Root–Pattern Data and Distribution in Arabic Morphology" MDPI 2018, 3, 10; doi: 10.3390/data3020010.
6. Selim, M. and Dan, A. **2007**. Simovici, "Bit Sequences and Biclustering of Text Documents," Seventh IEEE International Conference on Data Mining, DOI 0.1109/ICDMW.
7. Hung, C. and Xiaotie, D. **2008**. "Efficient Phrase-Based Document Similarity for Clustering," *IEEE Transactions On Knowledge And Data Engineering*, 20(9), September.
8. Mohammed, A. and Ophir, F. **2002**. "On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming," Approach ACM Eleventh Conference on Information and Knowledge Management (CIKM), McLean, VA, November, pp. 340-347.
9. Shreen, K., Porger, G. and Gerry, K. **2001**. "A tagset for the morphosyntactic tagging of Arabic," Paper presented at Corpus Linguistics 2001, Lancaster University, Lancaster, UK,.
10. Al-Fedaghi, S. and Al-Anzi, F. **1989**. "A new algorithm to generate Arabic root-pattern forms,". In proceedings of the 11th national Computer Conference and Exhibition. PP 391-400.
11. Anne, N., DE ROECK, Waleed AL-FARES **2001**. "A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots," Kuwait university.
12. Beesley, K. **1998**. "Arabic Morphological Analysis on the Internet," In the proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing, Cambridge, 17-18.
13. Abderrahim, B., Rachid, B., Abdelhak, L., Azzeddine, M., Abdelouafi, M. and Mohamed, B. **2011**. "A Markovian Approach for Arabic Root Extraction," *The International Arab Journal of Information Technology*, **8**(1).
14. Dilekh, T. and Behloul, A. **2012**. "Implementation of a New Hybrid Method for Stemming of Arabic Text", *International Journal of Computer Applications*, **46**(8): 14-19.
15. Alzamil, A. **2016** "Relation between roots and patterns," <http://ceri.kacst.edu.sa>. last enter 2016.
16. Khoja, S. **1999** "Stemming Arabic Text," Lancaster, U.K., Computing Department, Lancaster University. Available online at URL:<http://zeus.cs.pacificu.edu/shereen/research.thml>.