Corpus for Test, Compare and Enhance Arabic Root Extraction Algorithms

Nisrean Thalji
School of Computer and Communication Engineering
University Malaysia Perlis
Perlis, Malaysia

Yasmin Yacob
School of Computer and Communication Engineering
University Malaysia Perlis
Perlis, Malaysia

Nik Adilah Hanin
School of Computer and Communication Engineering
University Malaysia Perlis
Perlis, Malaysia

Sohair Al-Hakeem
Computer Science Department
University of Wales
Wales, UK

***Abstract*—Many studies have focused recently on building, evaluating and comparing Arabic root extracting algorithm. The main challenges facing root extraction algorithms are the absence of standard data set for testing, comparing and enhancing different Arabic root extraction algorithms. In addition, the absence of complete lists of roots prefixes suffixes and patterns. In this paper, we describe the development of a new corpus driven from traditional Arabic dictionaries "mu'jams". The goal is to use the corpus, as a new gold standard data set for testing, comparing and enhancing different Arabic root extraction algorithms. This data set covers all types of words and all roots. It contains each word and its root as a pair to avoid the consultation of a human expert needed to verify the correct roots of words used in the testing or comparing process. We describe the individual phases of the corpus construction, i.e. normalisation, reading derivation words and roots as a pair, and reading each root and its definition part. We have automatically extracted (12000) roots, (430) prefixes, (320) suffixes, (4320) patterns, and (720,000) word-root pair. Konja's and Garside Arabic root extraction algorithm was tested on this corpus; the accuracy was (63%), then we test it after supplying it with our lists of roots prefixes suffixes and patterns, the accuracy of it became 84%.**

***Keywords*—Arabic root extraction algorithm; corpus; pattern; prefix; suffix; root**

I. INTRODUCTION

Most researchers working in the field of Arabic root extraction algorithms opt to construct their own manually collected data set to run their experiments. Most of the time, the data sets are either small or incomprehensive. Therefore, their experimental findings may neither be convincing nor clear as for how to scale up the results [1].

The literature abounds with discussions about the design of Arabic stemming algorithms; yet little effort has gone into the investigation of the nature of the data set at the core of all these systems.

Al-Kabi and Al-Mustafa in [2], Ghwanmeh et el in [3], Al-Kabi et al in [4], Taghva et al in [5], Alshalabi in [6], Al-Shalabi and Evens in [7], Yaseen and Hmeidi in [8], Hmeidi et al in [9] and most new Arabic root extraction algorithms in the literature have tested their proposed root extraction algorithm on a different data set and compared their finding with other existing work. However, the data set that they used did not cover all types of words. In addition, the consultation by an Arabic language expert was needed to verify the accuracy of each finding manually.

Most of these algorithms manually constructed their own lists of prefixes, suffixes, and patterns as no standard lists were available. Thus, there was a huge variation between one algorithm and another. As the larger, the lists are the more accurate the result is.

Many research projects have studied Arabic root extraction algorithms and their effectiveness. Most of these studies claim an accuracy exceeding 75%. It has been found that the accuracy of these algorithms has been decreased after testing these algorithms on deferent data set other than what the researcher has used.

For example, in [3] Ghwanmeh et el claimed 95% accuracy for his algorithm. Testing the same algorithm in [4] on a different data set the authors claimed an accuracy of 67.40% for Ghwanmeh et el algorithm. Moreover, in [10] the authors conducted another test on Ghwanmeh et el algorithm using different data set. The author claimed an accuracy of 39%. This is due to a variation in size and type of the data set used to test Ghwanmeh et el stemmers [4].

As mentioned earlier, the lack of a standard data set was the main problem faced these algorithms. Each algorithm uses its own data set. These data sets are differed in size and type of words and are not available for authors to use.

Arabic root extraction algorithms need a standard data set to test their accuracy in comparison with other algorithms; this data set should be large enough to cover all types of words and cover all roots. This data set should contain the word and its root as a pair. In addition, Arabic root extraction algorithms need complete lists of roots, prefixes, suffixes, and patterns to enhance their accuracy.

The quality and coverage of the data set will determine the quality and coverage of each Arabic root extraction algorithm, and any limitations found in the data set will make their way through to the algorithm.

Arabic root extraction is an important step toward conducting effective research on most of the Arabic natural language processing (ANLP) applications.

Arabic root extraction algorithms are used in information retrieval systems, indexers, text mining, text classifiers, data compression, spelling checkers, text summarisation and machine translation. The algorithms extract stems or roots of different words, so that words derived from the same stem or root are grouped together.

In Latin-based languages, the stem and the root are the same; however, this is not the case for the Arabic language. Stemming is the first step toward finding the root. The stem is simply defined as a word without a prefix or/and suffix [11]. Some further processing to a stem through the removal of some infixes might be required to obtain an Arabic root.

For example, the stem from the word "القادمون" is "قادم", where the root is "قدم" [11].

The lack of a gold standard dataset to be used to carry benchmark tests of different Arabic root extraction algorithms lead us to develop and build an automated corpus (Gold standard dataset). The purpose of this corpus is to be used to test, compare and enhance different Arabic root extraction algorithms.

The standard gold data set:

- Should be large enough to contain all types of words and roots. There exist about 12000 roots.
- The data set should contain the word and its root to avoid the interference of a human expert normally needed to verify the correct roots of each word used in the testing or comparison process.

Our aim in this paper is to build a corpus pairing each word to its root and contain a standard list of roots, prefixes, suffixes, and patterns. The suggested corpus will help researchers to enhance, test and compare the present root-extraction algorithms and any future algorithms.

The structure of this paper is as follows. In Section 2, previous approaches and their drawbacks have been discussed. Section 3 describes proposed methodology, including details of each process. Section 4 explains the experimental implementation of our approach and the evaluation process. Section 5 concludes the main points of the paper and gives some future directions.

II. Previous Works

Khoja and Garside in [12] build corpus for the purpose of Arabic root extraction , which contains (7) diacritic characters, (38) punctuation characters, (5) definite articles, (168) stop words, (11) prefixes, (28) suffixes, (3,822) trilateral roots, (926) quadrilateral roots and (46) trilateral root patterns.

The corpus exists freely and publicly for researchers to download. The main issue here is that Khoja's corpus is limited in its contents, manually tagged and missing roots derivatives.

Buckwalter in [13] build corpus for purpose of Arabic morphological analyser, which contains (299) prefixes, (618) suffixes, (4,749) roots including both trilateral and quadrilateral roots, (82,185) stems, (38,600) lemmas, (1,648) prefix-stem combinations, (1,285) stem-suffix combinations and (598) prefix-suffix combinations.

Al-Shawakfa et al in [10] builds a corpus for the purpose of evaluating and comparing Arabic root extraction algorithms. This corpus was built based upon the set of trilateral Arabic roots that were introduced by Buckwalter in [13].

The developed corpus was mainly built of 3823 trilateral roots. By using these roots as a base, a corpus was obtained of approximately 27.6 million unique words of size 1. 63GB. Furthermore, all combinations of 73 trilateral patterns, 10 suffixes, and eight prefixes were applied to the roots to create different forms of Arabic words. All generated words were syntactically correct; but not necessary semantically correct.

Al-Shawakfa corpus did not require a manual root verification upon completing the testing process.

The disadvantages of Al-Shawakfa corpus are:

- In many cases, many words are not semantically correct.
- Although the fact that the corpus has contained large data set, it has only covered 3823 roots out of 12000.
- Two types of words are missing:

1) Words with (changing the vowel letters with deferent vowel letters "الاقلاب ".For example, the root "قول", "و" letter is changing to "ا" in "قال" word.

2) Words with (changing the place of a letter " الابدال ") type. For example, the root "وجه", "و" letter is changing to "ا" in "جاه"word, and the place of "ا" has changed in the new word too.

Sawlha and Atwell [14] constructed a broad-coverage lexical resource to improve the accuracy of morphological analysers and part-of-speech taggers of Arabic text. Twenty-three lexicons have been collected from different web resources freely available.

The lexicons' texts contain 14,369,570 words, 2,184,315 vowelised word types and 569,412 non-vowelised word types. According to Sawalha and Atwell's study, a tokenising module for the program must specify the root entries and their definition parts. Then, a bag of words is extracted from the definition text. The bag stores pairs of word-root where each word appearing on the definition part is associated to the root of that part.

Many words appearing in the definition part are not relevant to the root associated with that definition. Such words are found inside the bag of words- root. A normalisation

analysis that verifies the word-root pairs is done by applying linguistic knowledge that governs the derivation process of words from their roots. These conditions are simply described as the following:

Condition 1 (check consonants): If all consonant letters constructing the root appear in the analysed word, then check condition 2.

Condition 2 (consonants order): If all root letters appear in the same order as the word's letters, then word-root combination might be correct. [14]

Since the Arabic language is a sophisticated language, these two conditions are not enough to be sure that this word is derived from this root. Sawalha and Atwell algorithm was implemented. The algorithm has retained successes in some cases and fails in many cases.

Sawalha and Atwll research is a step forward towards creating a new corpus derived from Arabic lexicons to be used as a standard data set containing all the roots, a large number of derivatives and pairing each root with its derivatives. Our finding shows that there are many words are related to unexpected roots.

Table 1 shows an example of words that are wrongly related to the roots. In addition, the algorithm doesn't declare how many pairs of words roots were founded. It is clear; this work needs more rules to enhance the results.

TABLE I. EXAMPLE OF WORDS THAT ARE WRONGLY RELATED TO THE ROOTS BY SAWALHA AND ATWELL CORPUS

No	Word	Root	Root by Sawalha and Atwell
1	والسَّبَّابةِ	سبب	ألب
2	والسَّهْبُ	سهب	أوب
3	نفسه	نفس	أفف
4	الناس	نوس	أوأ
5	النَّباتِ	نبت	أبب
6	الأسماء	سما	أيا
7	وتأملنا	أمل	أيا
8	منافعِ	نفي	أما
9	ثوبٌ	ثوب	ثأي
10	حاجبِ	حجب	حجج
11	الفٌ	ألف	أوأ
12	أهله	أهل	أول
13	المِنْشارُ	نشر	أشر
14	أطلب	طلب	أبب

III. METHOD

All Arabic roots and its derivations can be found in ("mu'jams","المعاجم") dictionaries. Most of the Arabic dictionaries were studied carefully in this paper.

Traditional Arabic lexicons are not available in computerised lexicographic databases. Moreover, they have different arrangement methodologies than modern English dictionaries [14]. Existing Arabic dictionaries suffer from many issues. The main one was that they were built to be used manually. Dictionaries in Arabic contain the roots as a title followed by root definition part, which may contain one or more paragraphs for each root; these paragraphs describe the meaning of the root and contain possible word's derivation from the root. The definition part may extend to many pages. Each dictionary has its own deferent definition part. New and deferent information can be read for each root when reading different dictionaries. Figure 1 shows a sample of text taken from Al-Mesbah-Almonir dictionary ("معجم المصباح المنير"), with roots ("أبد, ابد, أبر") and its definition parts. Figure 2 shows a sample of text taken from Asas Al-Blaghah dictionary ("معجم أساس البلاغة"), with roots ("أبد, ابد, أبر") and its definition parts. We can notice the deferent information given each time.

[أ ب ب] الأَبُ : المَرْعَى الَّذِي لَمْ يَزْرَعْهُ النَّاسُ مِمَّا تأْكُله الدَّوابُّ والْأَنْعَامُ وَيُقَالُ (الْفَاكِهَةُ لِلنَّاسِ والْأَبُ لِلدَّوابِّ) وقالَ ابنُ فَارِسٍ قالُوا (أَبَ) الرَّجُلُ (يَؤُبُ) (أَبْاً وأَبَاباً وأَبَابَةً) بِالْفَتْحِ إِذا تَهَيَّأَ لِلذَّهابِ ومِنْ هُنا قِيلَ (الثَّمَرةُ الرَّطْبةُ هِيَ الْفَاكِهةُ والْيَابِسُ مِنْهَا الأَبُ) لأَنَّهُ يُعَدُّ زاداً لِلشِّتاءِ والسَّفَرِ فَجُعِلَ أَصْلُ الأَبِ الاسْتِعْدَادَ و (الإِبَّانُ) بِكَسْرِ الْهَمْزَةِ والتَّشْدِيدِ الْوَقْتُ إنما يُسْتَعْمَلُ مُضَافاً فَيُقَالُ (إِبَّانُ) الفاكِهةِ أَيْ أَوَانُها ووَقْتُهَا ونُونُهُ زائِدَةٌ مِنْ وَجْهٍ فَوَزْنُهُ فِعْلانٌ وأَصْلِيَّةٌ من وَجْهٍ فوزنُه فِعَّالٌ.

[أ ب د] الأَبَدُ : الدَّهْرُ ويُقَالُ الدَّهْرُ الطويلُ الَّذي لَيْسَ بمحدُودٍ قال الرُّمَّانيُّ فإِذَا قلتَ لا أُكَلِّمُهُ (أَبَداً) فالأَبَدُ من لَدُنْ تَكَلَّمْتَ إِلى آخِرِ عُمْرِكَ وجَمْعُهُ (آبَادٌ) مِثْلُ سبَبٍ وأَسْبَابٍ و (أَبَدَ) الشَّيءُ مِنْ بَابَيْ ضَرَبَ وقَتَل (يَأْبِدُ) و (يَأْبُدُ) (أُبُوداً) نَفَر وتوحَّشَ فَهُوَ (آبِدٌ) عَلَى فَاعِلٍ و (أَبَدَتِ) الوُحوشُ نَفَرتْ من الإِنسِ فَهِيَ (أَوابِدُ) ومِنْ هُنَا وُصِفَ الْفَرَسُ الخفيفُ الذى يُدْرِكُ الوَحْشَ ولا يَكَادُ يَفُوتُه بِأَنَّهُ (قَيْدُ الأَوَابِدِ) لأَنَّهُ يَمْنَعُهَا المُضِيَّ والخَلاصَ مِنَ الطَّالِبِ كَمَا يَمْنَعُهَا القَيْدُ وقِيلَ لِلأَلْفَاظِ الَّتِي يَدِقُّ معناها (أَوابِدُ) لِبُعْدِ وُضُوحِهِ لأَنَّه المقْصُودُ.

[أ ب ر] أَبَرْتُ : النَّخْلَ (أَبْراً) من بَابَيْ ضرَب وقَتل لقَحْتُه (وأَبَّرْتُهُ) (تَأْبِيراً) مُبَالَغَةٌ وتَكْثِيرٌ (وَالْأَبُور) وزَانُ رَسُولٍ ما يُؤَبَّر بِهِ (وَالإِبَارُ) وزَانُ كِتَابٍ النَّخْلُةُ الْأُنثى (يُؤَبَّر) بطَلْعِهَا وقِيل (الإِبَارُ) أيضاً مَصْدرٌ كَالقِيام والصِّيَامِ و (تَأَبَّرَ) النخلُ قَبِل أن (يُؤَبَّر) قال أبو حاتمٍ السِّجِسْتَانِيُّ فى كتاب النخلة إذا انْشَقَّ الكَافُورُ قِيلَ شَقَق النَّخْلُ وَهُوَ حِينَ (يُؤَبَّرُ) بالذَّكَرِ فَيُوتَى بشَمَارِيخِهِ فَتُنْفَضُ فيَطِيرُ غُبارُها وهُوَ طَحِينُ شَمَارِيخِ الفَحَّالِ إِلىَ شَمَارِيخِ الأُنْثى وذلِكَ هُوَ التَّأْبِيحُ و (الْإِبْرَةُ) مَعْرُوفةٌ وَهِيَ الْمِخْيَطُ والْخِيَاطُ أَيْضاً والْجَمعُ (إِبَرٌ) مِثْلُ سِدْرةٍ وسِدَرٍ.

Fig. 1. Sample text of Al-Mesbah-Almonir dictionary.

The definition part is written as an article which defines most of the derived words of a certain root and contains many other words. These words are neither the root nor its derivatives. They exist mainly for explaining the meaning of the root. In Figure 1 the roots are written between two brackets; the derived words are written between two parentheses with red colour. This is a modified version of the original dictionary; the original version did not distinguish between the roots and its derivation.

The problem of the modified version is that many parentheses contain words other than the root derivation words. In addition, not all the root derivation words are written between two parentheses.

Unfortunately, most of the existing Arabic dictionaries do not distinguish between the roots and its derivation. In Figure 2 the roots are written on separate lines followed by their definitions. Again, in many other places in the dictionary you

will find other words are written on a separate line, and these words are not roots. In other places in the dictionary, the roots are written at the beginning of the paragraphs. These dictionaries are written without any computerised lexicographic representations. Manual work was carried out to distinguish the roots from other entries.

أبب
اطلب الأمر في إبانه، وخذه بربانه، أي أوله، وأنشد ابنُ الأعرابي:
قد هرمتني قبل إبان الهرم ... وهي إذا قلت كلي قالت نعم
صحيحة المعدة من كل سقم ... لو أكلت فيلين لم تخش البشم
وأبَّ للمسير إذا تهيّأ له وتجهز. قال الأعشى: صرمت ولم أصرمكم وكصارم ... أخ قد طوى كشحاً وأب ليذهبا ونقول: فلان راع له الحب، وطاع له الأب، أي زكا زرعه واتسع مرعاه.
أبد
لا أفعله أبد الآباد، وأبد الأبيد، وأبد الآبدين. ونقول: رزقك الله عمراً طويل الآباد، بعيد الآماد، وأبدت الدواب وتأبدت: توحشت، وهي أوابد ومتأبدات. وفرس قيد الأوابد وهي نفر الوحوش. وقد تأبد المنزل: سكنته الأوابد. وتأبد فلان: توحش. وطيور أوابد خلاف القواطع. ومن المجاز: فلان مولع بأوابد الكلام وهي غرائبه، وبأوابد الشعر وهي التي لا تشاكل جودة. قال الفرزدق:
لن تدركوا كرمي بلؤم أبيكم ... وأوابدي بتنحل الأشعار
وقال النابغة: نبئت زرعة والسفاهة كاسمها ... يهدي إليّ أوابد الأشعار وجئتنا بأبدة ما نعرفها.
أبر
شاة مأبورة: أكلت الإبرة في علفها. وعن مالك بن دينار مثل المؤمن كمثل الشاة المأبورة. ويقال: أشد من وخز الإبر. وأبر النخل وأبره. وتأبر النخل: قبل الإبار. وتقول: إذا رفق الأبار، سحق الجبار. ومن المجاز: إبرة القرن لطرفه. قال ابن الرقاع:
تزجي أغن كأن إبرة روقه ... قلم أصاب من الدواة مدادها
وإبرة المرفق لطرفه، وإبرة العقرب والنحلة لشوكتها. وتقول: لا بدّ مع الرطب من سلاء النخل، ومع العسل من إبر النحل. وقد أبرته العقرب بمئبرها والجمع مآبر. ومنه: إنه لذو مآبر في الناس كما قالوا: دبت بينهم العقارب إذا مشت بينهم النمائم. وقال النابغة: وذلك من قول أتاك أقوله ... ومن دس أعداء إليك المآبرا وأبرني فلان إذا اغتابك وآذاك. وتقول: خبئت منهم المخابر، فمشت بينهم المآبر.

Fig. 2. Sample of text taken from Asas Al-Blaghah dictionary

Our study takes the following traditional Arabic lexicons:-

“Kitab Al-'Ayn” by Al-Khalil Al-Farahidi in [15], “Lisan Al-Arab“ by Ibn Manzur in [16], “Tag Al-‘Arus Min Gawahir Al-Qamus” by Al-Zabidi in [17], “Asas Al-Balaghah” by Abu-Al-Qasim Maḥmud Bin ‘Amr Bin Aḥmad Al-Zamahshari in [18], “Al-Mugrib Fi Tartib Al-Mu‘Rib” by Abu Al-Fatḥ Naṣir Ad-Din Al-Mutrazi in [19], “Mukhtar As-Sihah” by Abu Bakr Al-Razi in [20], “Al-Musbah Al-Munir Fi Garib Al-Sharh Al-Kabir” by Aḥmad Bin Muḥammad ‘Ali Al-Fayyumi in [21], “Al-Muḥiṭ Fi Al- Luga” by Abu Al-Qasem Al-Ṣaḥib Bin ‘Abbad in [22], “Al-Ṣiḥaḥ Fi Al-Luga” by Abu Naṣr ’Isma‘il Bin Ḥammad Al-Gawhari Al-Farabi in [23], and finally “Kalemat Al-Quraan Al-kaream” by mohammed kheder in [24].

A. Manual Annotations

Traditionally, lexicons are constructed in many ways. Roots and lexical entries are presented without using any computerised lexicographic representations, and the roots of many of them are not distinguishable from other entries.

In this study, the root has been distinguished manually from other entries. Each root has been placed between two stars symbol “*”. Figure 3 shows a sample text of Asas Al-Balaghah dictionary after putting each root between two stars. The process has covered all existing traditional dictionaries to enable the researchers from reading each root and its definition part automatically.

أبب
اطلب الأمر في إبانه، وخذه بربانه، أي أوله، وأنشد ابنُ الأعرابي:
قد هرمتني قبل إبان الهرم ... وهي إذا قلت كلي قالت نعم
صحيحة المعدة من كل سقم ... لو أكلت فيلين لم تخش البشم
وأبَّ للمسير إذا تهيّأ له وتجهز. قال الأعشى: صرمت ولم أصرمكم وكصارم ... أخ قد طوى كشحاً وأب ليذهبا ونقول: فلان راع له الحب، وطاع له الأب، أي زكا زرعه واتسع مرعاه.
أبد
لا أفعله أبد الآباد، وأبد الأبيد، وأبد الآبدين. ونقول: رزقك الله عمراً طويل الآباد، بعيد الآماد، وأبدت الدواب وتأبدت: توحشت، وهي أوابد ومتأبدات. وفرس قيد الأوابد وهي نفر الوحوش. وقد تأبد المنزل: سكنته الأوابد. وتأبد فلان: توحش. وطيور أوابد خلاف القواطع. ومن المجاز: فلان مولع بأوابد الكلام وهي غرائبه، وبأوابد الشعر وهي التي لا تشاكل جودة. قال الفرزدق:
لن تدركوا كرمي بلؤم أبيكم ... وأوابدي بتنحل الأشعار
وقال النابغة: نبئت زرعة والسفاهة كاسمها ... يهدي إليّ أوابد الأشعار وجئتنا بأبدة ما نعرفها.
أبر
شاة مأبورة: أكلت الإبرة في علفها. وعن مالك بن دينار مثل المؤمن كمثل الشاة المأبورة. ويقال: أشد من وخز الإبر. وأبر النخل وأبره. وتأبر النخل: قبل الإبار. وتقول: إذا رفق الأبار، سحق الجبار. ومن المجاز: إبرة القرن لطرفه. قال ابن الرقاع:
تزجي أغن كأن إبرة روقه ... قلم أصاب من الدواة مدادها
وإبرة المرفق لطرفه، وإبرة العقرب والنحلة لشوكتها. وتقول: لا بدّ مع الرطب من سلاء النخل، ومع العسل من إبر النحل. وقد أبرته العقرب بمئبرها والجمع مآبر. ومنه: إنه لذو مآبر في الناس كما قالوا: دبت بينهم العقارب إذا مشت بينهم النمائم. وقال النابغة: وذلك من قول أتاك أقوله ... ومن دس أعداء إليك المآبرا وأبرني فلان إذا اغتابك وآذاك. وتقول: خبئت منهم المخابر، فمشت بينهم المآبر.

Fig. 3. Sample text of Asas Al-Balaghah dictionary after distinguishing the roots

B. Normalisation

Text normalisation is defined as a process that consists of a series of steps that should be followed to wrangle, clean and standardise textual data to a form which could be consumed by other NLP and analytics systems and applications as input [13].

The process steps of the proposed text normalisation are as follows:

1) Remove kasheeda symbol ("_").
2) Remove punctuations.
3) Remove diacritics.
4) Remove non-letters.
5) Replace hamza’s forms ء , آ, ؤ, إ, ئ with أ .
6) Duplicating any letter that has the (Shaddah " ّ ") symbol.

C. Extract All Information

In this section, we try to read all information in dictionaries.

1) Extract Roots and Their Definitions Part

A separate database was created and saved for each studied dictionary. The created database consists of the distinguished root and their definition part. Table 2 shows a sample of the created database for some roots and their definitions parts taken from Asas Al-Balaghah dictionary.

TABLE II. SAMPLE OF THE DATABASE FOR ROOTS AND THEIR DEFINITIONS FOR ASAS AL-BALAGHAH DICTIONARY

Full Text	Root
اطّلب الأمر في إبانه وخذه بربانه أي أوله وأنشد ابنُ الأعرابي قد هرمتني قبل إبان الهرم وهي إذا قلت كلي قالت نعم صحيحة المعدة من كل سقم لو أكلت فيلين لم تخش البشم وأب للمسير إذا تهيّأ له وتجهز قال الأعشى صرمت ولم أصرمكم وكصارم أخ قد طوى كشحاً وأب ليذهبا ونقول فلان راع له الحب وطاع له الأب أي زكا زرعه واتسع مرعاه	أبب
لا أفعله أبد الآباد وأبد الأبيد وأبد الأبدين ونقول رزقك الله عمراً طويل الآباد بعيد الآماد وأبدت الدواب وتأبدت توحشت وهي أوابد ومتأبدات وفرس قيد الأوابد وهي نفر الوحوش وقد تأبد المنزل سكنته الأوابد وتأبد فلان توحش وطويل أوابد خلاف القواطع ومن المجاز فلان مولع بأوابد الكلام وهي غرائبه وبأوابد الشعر وهي التي لا تشاكل جودة قال الفرزدقلن تدركوا كرمي بلؤم أبيكم وأوابدي بنتحل الأشعار وقال النابغة نبئت زرعة والسفاهة كاسمها يهدي إليّ أوابد الأشعار وجئتنا بأبدة ما نعرفها	أبد
شاة مأبورة أكلت الإبرة في علفها وعن مالك بن دينار مثل المؤمن كمثل الشاة المأبورة ويقال أشد من وخز الإبر وأبر النخل وأبره وتأبر النخل قبل الإبار وتقول إذا رفق الأبار سحق الجبار ومن المجاز إبرة القرن لطرفه قال ابن الرقاع تزجي أغن كأن إبرة روقه قلم أصاب من الدواة مدادها وإبرة المرفق لطرفه وإبرة العقرب والنحلة لشوكتها وتقول لا بدّ مع الرطب من سلاء النخل ومع العسل من إبر النحل وقد أبرته العقرب بمئبرها والجمع مآبر ومنه إنه لذو مآبر في الناس كما قالوا دبت بينهم العقارب إذا مشت بينهم النمائم وقال النابغة وذلك من قول أتاك أقوله ومن دس أعداء إليك المآبرا وأبرني فلان إذا اغتابك وآذاك وتقول خبئت منهم المخابر فمشت بينهم المآبر	أبر

2) Extract Derivation Words and Roots As Piar

Using the All the derivation words of each root are extracted from the definition parts using the following algorithm:

(Condition 1), examine consonants in the root:

If all consonants letters constructing the root appeared in the analysed word, then continue, else the word is rejected, consonants letters are all the letters except vowels. For example, in the root "كتب" the word that doesn't contain "ك" , "ت" or "ب" letters is rejected. For example, the word "كتبوا" , is accepted and the word "تابوا" is rejected , so the pair(, "كتبوا" "كتب") is accepted, and the pair ("كتب" , "تابوا ") is rejected.

(Condition 2), examine consonants order:

If all root's letters appear in the same order as the word's letters, then continue, else the word is rejected.

For example, in the root "كتب" if any of the words contain "ك" , "ت" or "ب" letters in deferent order than it appears in the root, the word will be rejected. This implies that the pair ("بكت" "كتب") is rejected.

(Condition 3), examine consonants in the word:

If the word contains at least one of these letters," 'ث', 'ج', 'ح', 'خ', 'د', 'ذ', 'ر', 'ز', 'ش', 'ص', 'ض', 'ط', 'ظ', 'ع','غ', 'ق' " and these letters are not in the root then the word is rejected. For example, in the root "حرج" the word "دحرج" is rejected. The word contains the letter "د" which doesn't exist in root letters, so the pair (" حرج " , "دحرج") is rejected.

(Condition 3.1), examine replacing letter by another in the word"الإبدال":

There are exceptions for condition 3 for "ط, د" letters, if "د" comes after "ز" letter in the word, "د" may be omitted from the root, for example, the word "ازدهر" is accepted for "زهر" root, so the pair ("زهر" , "ازدهر") is accepted, in some cases "د" is converted to "ت"for simplifying purpose.

if "ط" comes after "ض, ص, ط , ظ" in the word, "ط"may be omitted from the root, for example, the word "اصطبر" is accepted for "صبر" root, so the pair (" صبر " , "اصطبر") is accepted, in some cases "ط" is converted to "ت" for simplifying purpose.

(Condition 4), examine vowels in the root:

If the root contains vowel's letters "ي , و , ا", or Hamza "أ" it's not necessary to accept the root that contains an exact letter. Words that contains vowel letters is accepted for the same root following the rule of Ebdal "ابدال" in Arabic. For example, in the root "قول"we accept the word "قال", so the pair ("قال" , " قول") is accepted.

(Condition 5), examine vowels order:

If all root's vowel letters appear in the same order as the word's vowel letters, then the word is accepted. For example, in the root "قول" and word "قال" the vowels are in same order, which is after "ق" letter and before "ل" letter, so the pair (, "قال" "قول ") is accepted.

If some or all of the vowel's letter in the root are appearing in deferent order, the word is a candidate but not sure true. Like the word "عدو" is not derived from the root "وعد". In this case the word is wrongly related to the root, but in other cases is not, like the word "بيغ" and the root "بغي". These candidate words are examined in all other dictionaries, if the root is the same, we change it to be true. So initially these pairs are rejected until we compare it with other dictionaries.

(Condition 6), examine the existence of vowels:

If some or all of vowel's letters in the root are not appearing in the word, the word is considered as a candidate but not sure true. For example, the word "قل" and the root "قول". In this example, the word is truly related to the root, but like the word "أن" it is not derived from the root "أين". These candidate words are examined in all other dictionaries, if the root is the same, the root will be changed to a candidate root. Initially these pairs are rejected until it has been compared with other dictionaries.

(Condition 7), examine root for duplication letter

If the root has a duplication letter like the root "بجج", the word that has one or two "ج" letter will be accepted, like "بجت" word and "بججت", so these pairs are accepted,("بجت" ,"بججج") and ("بجج" ,"بججت").

Furthermore, for this type of root the word that repeats the full root letters after the first full root letters will be accepted, for example, the word "بجبجت" for the root "بجج" , so the pair ("بجج" ,"بجبجت") is accepted.

(Condition 8), examine these rules:

Our work is opposite to the root extraction algorithms work, they start from the derivation words to find the root. In our work, the root is known initially and then the derivation words have to be found. When the root is known, finding the derivation words is much easier than finding the root. We have used the rules that were discovered by root extraction algorithm in [25]. These rules are mention below:

Prefix letters:

These letters can be added only in the prefix part. They are :{ل ،س، ف، ب}. Prefix part: is the part of the word, one or more letters before the first letter of the root word. So if we found these letters in place other than prefix part, and these letters are not a root's letter this word will be rejected. In root extraction algorithm finding, the prefix part is a challenge, but in our work, we can determine the prefix part as the root is known. For example, the word "السبابة" founded in the definition part of "سبب" root, so "س , ب , ب" letters are consonant in the word, the part before "س" letter in the word is the prefix, which is "ال". This word is accepted because"ل" can be in prefix part, so the pair ("السبابة, سبب") is accepted. And "السبابة" also founded in the definition part of "ألب" root, so "ب , ب , ا" letters are consonant in the word, but "س" is not in prefix place, its after "ا", so the pair ("السبابة, ألب") is rejected.

Suffix letters:

These letters can be added only in the suffix part. Suffix part: Is the part of the word, one or more letters after the last letter of the original root word. So if we found these letters in place other than suffix part and these letters are not root's letters the word is rejected. In this paper the Suffixes are limited to single letter suffix: {ه}.

For example, the word "أوله" founded in definition part of "أول" root, so "أ , و , ل" letters are consonant, the part after "ل" letter in the word is the suffix part, which is "ه". This word is accepted as " ه" was founded in the suffix part, so the pair ("أوله, أول") is accepted.

Another example is the word "أهوي" founded in the definition part of "أوي" root, so "أ , و , ي" letters are consonant, "ه" is not a root letter, and has not been found in the suffix part, it is before "ي" and not after, so the pair ("أهوي, أوي") is rejected.

Prefix-Suffix letters:

These letters can be found only on both sides of the word, i.e. in the suffix part or in the prefix part. They are :{م, ك , ن}.

If these letters have been spotted in place other than prefix part or suffix part in the word, and these letters are not a root's letters this word is rejected. For example, the word "أنبض" and the root "أبض", this word is rejected because " ن" is not in prefix or suffix places, it's neither before "أ" nor after "ض", so the pair ("أنبض, أبض ") is rejected.

All roots and their derivation words are stored in a database. Table 3 shows a sample from the database for " Asas Al-Balaghah" dictionary after picking the derivation words.

The database contains the roots, their derivation words, and the definition part for each root; derivation words were distinguished by putting each derivation word between brackets.

TABLE III. SAMPLE OF THE DATABASE FOR " ASAS AL-BALAGAH" AFTER PICKING THE DERIVATION WORDS.

Root	Derivatives Words	Definition Part
حشد	حشد حشوداً واحتشدوا وتحشدوا وتحاشدوا وحشدتهم أحشدهم وأحشدهم حشداً حشد محشود واحتشدت واحتشد واحتشد حاشد والحاشدون حاشد تحشد	[حشد] القوم [حشودا] و اجتمعوا وخفوا في التعاون [واحتشدوا] [وتحشدوا] [وتحاشدوا] على الأمر اجتمعوا عليه متعاونين [وحشدتهم] [أحشدهم] [وأحشدهم] حشداً وعنده [حشد] من الناس ورجل [محشود] محفود مجتمع عليه مخدوم [واحتشدت] لفلان في كذا أعددت له [واحتشد] لنا في الضيافة إذا اجتهد وبذل وسعه [واحتشد] للضيافة احتفل لها وفلان حافد [حاشد] مجتهد في خدمته وضيافته وسعيه قال [والحاشدون] على قرى الأضياف وإذا كان للإبل من يقوم بحلها لا يفتر عنه قالوا لها حالب [حاشد] ومن المجاز بت في ليلة [تحشد] عليّ الهموم.

3) Extract Prefixes, Suffixes and Patterns

Since the root and its derivation words are known, prefix, suffix and the pattern can be extracted from each word. "ف" will replace the first root letter, "ع" will replace second root letter, and "ل" will replace the third root letter in the word. If the root is more than three letter length, "ل" will replace all the rest of the letters.

For example in Table 3 the trilateral root is "حشد" and the derivation words are:

"حشد, حشوداً, واحتشدوا ,وتحشدوا, وتحاشدوا, وحشدتهم, أحشدهم وأحشدهم ,حشداً, حشد, محشود, واحتشدت, واحتشد, واحتشد ,حاشد والحاشدون, حاشد, تحشد ".

"ف" will replace "ح","ع" will replace "ش", and "ل" will replace "د" in all derivation words for "حشد" root. Now the patterns are:-

"فعل, فعولاً, وافتعلوا ,وتفعلوا, وتفاعلوا, وفعلتهم, أفعلهم وأفعلهم ,فعلاً, فعل, مفعول, وافتعلت, وافتعل, وافتعل ,فاعل والفاعلون, فاعل, تفعل ".

And for quadratic root like "دحرج", "ف" replace "د","ع" replace "ح", "ل" replace "ر", and "ل" replace "ج" in the derivation words for "دحرج" root. So the derivation words like("التدحرج, متدحرجون, دحرجوهم") the patterns will be ("التفعلل, متفعللون, فعللوهم").

The part of the pattern before "ف" is considered as prefix and the part of the pattern after last "ل" is considered as a suffix, so "فعل" has no prefix or suffix, "والفاعلون" has "وال" prefix and "ون" suffix. All prefixes suffixes and patterns in all dictionaries were collected saved in a separate database, Table 4 shows a sample of the database for prefixes suffixes and patterns.

TABLE IV. SAMPLE OF THE DATABASE FOR PREFIXES SUFFIXES AND PATTERNS

No	Word	Pattern	Prefix	Suffix
1	والمتصدقات	والمتفعلات	والمت	ات
2	متصدقين	متفعلين	مت	ين
3	أتحدثونهم	أتفعلونهم	أت	ونهم
4	فأخرجناهم	فأفعلناهم	فأ	ناهم
5	وتستخرجوا	وتستفعلوا	وتست	وا
6	وتستخرجون	وتستفعلون	وتست	ون
7	والخاشعات	والفاعلات	وال	ات
8	سنستدرجهم	سنستفعلهم	سنست	هم

Now our corpus contains (12000) roots, (430) prefixes, (4320) patterns, (720,000) word-root pair.

IV. EXPERIMENT AND EVALUATION

In this section a comparison between our corpus, Khoja and Garside corpus, Buckwalter corpus, and Al-Shawakfa et al corpus was conducted. The result of the comparison is shown in Table 5.

TABLE V. COMPARISON BETWEEN OUR CORPUS, KHOJA AND GARSIDE CORPUS, BUCKWALTER CORPUS, AND AL-SHAWAKFA ET AL CORPUS

Corpus	No of root	No of prefixes	No of suffixes	No of patterns	No of word root pair
Khoja and Garside corpus	4748	11	28	46	0
Buckwalter corpus	4,749	299	618	3531	0
Al-Shawakfa et al corpus	3823	8	10	73	276000000
Our corpus	12000	430	320	4320	720000

The Table 5 shows that Khaja and Buckwalt corpuses have not paired each word with its root. As mention earlier, Khojas corpus has limited number of suffixes, prefixes and patterns. It has been shown that Shawakfa corpus has more suffixes, Prefixes and pattern in comparison with Khoja's corpus. Our corpus has the longest lists of roots, prefixes, suffixes and patterns. Al-Shawakfa et al corpus have the longest list of the word-root pair, but as mention in previous work section many words are semantically incorrect.

Khoja and Garside reported 96% accuracy of her stemmer using newspaper text on the assumption it was evaluated on the developed corpus. However, details of the evaluation methodology are not available, the text used in evaluation and accuracy metrics[26].

Khoja and Garside algorithm was tested in many studies; it was tested in [10] study, the test reveals an accuracy of 34%, and tested in [3] study, the test reveals an accuracy of 74%. This is due to differences in size and type of the data sets that are used[4]. The main challenges or problems that faced authors wanted to test or compare these algorithms are the manual verification for a result, and the absence of a corpus that has the word and its root as a pair.

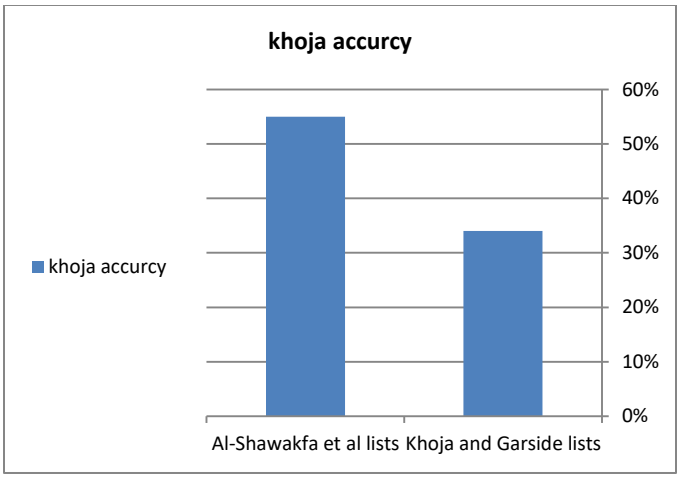

Fig. 4. Khoja and Garside algorithm's accuracy before and after supplying Al-Shawakfa et al corpus's lists

Khojas algorithm was tested using Al-Shawakfa corpus. An accuracy of 34% was obtained initially. The accuracy of the test has increased to 55% after providing Khoja's algorithm with Al-Shawakfa corpus lists, see Figure 4.

Khoja and Garside algorithm was tested on the newly developed corpus to compute the accuracy of their algorithm. Khoja and Garside Algorithm achieved about (63%) average accuracy. This is due to many factors:

Restricting the result for just (4748) roots, (3,822) trilateral roots, (926) quadrilateral roots. It has ignored (7252) roots, for example, the word "إبانه" is stemmed is to the wrong root "بين", because the root "أبب"is missing.

Missing a very large number of prefixes, suffix, and patterns, for example, the word "حوسب" is not stemmed, because it is missing the pattern "فوعل".

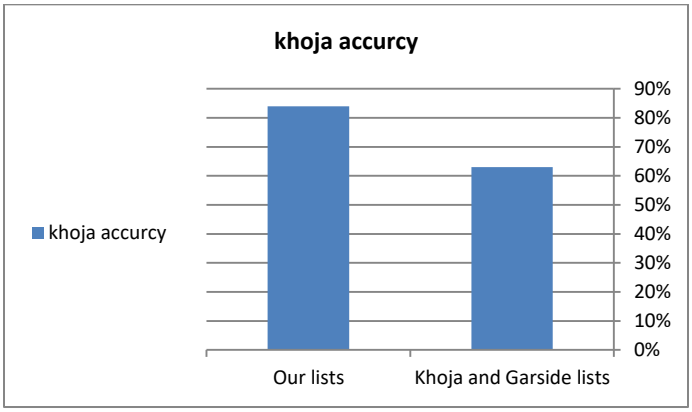

Fig. 5. Khoja and Garside algorithm's accuracy before and after supplying our corpus's lists

Another test was conducted on Khoja and Garside algorithm after supplying the newly developed corpus with our lists of roots, prefixes, suffixes, and patterns. Khoja and Garside algorithm has achieved (84%) average accuracy.

Figure 5 shows Khoja and Garside algorithm accuracy average rate before and after supplying the newly developed corpus's lists .

V. Conclusion And Future Work

In this work, a new corpus has been developed based on traditional manual Arabic dictionaries "mu'jams". The developed corpus was built mainly for testing, comparing and enhancing Arabic root extraction algorithms; we automatically extracted from these dictionaries (12000) roots, (430) prefixes, (320) suffixes, (4320) patterns, (720,000) word-root pair.

The developed corpus covers all types of words and all roots. It contains each word paired with its root. The developed corpus will save a lot of time and effort compared with the manual corpus previously used for testing purposes.

There is no need for the manual verification usually done by consulting Arabic language experts. Arabic root extraction algorithms can test and compare their finding using the newly automated corpus.

Khoja and Garside Arabic root extraction algorithm was tested using the developed corpus. The test has given results with (63%) accuracy.

The test was carried out after supplying it with our lists of roots prefixes, suffixes, and patterns the accuracy of it becomes 84%.

We plan to enhance the accuracy of Khoja and Garside algorithm and solve problems such as affix ambiguity, Ebdal and Eqlab, stop words, foreign words and the problem with one solution.

References

[1] B. Hammo, F. Al-Shargi, S. Yagi and N. Obeid, "Developing tools for Arabic corpus for researchers," Paper presented at the Second Workshop on Arabic Corpus Linguistics (WACL-2), 2013.

[2] M. N. Al-kabi and R. AL-Mustafa, "Arabic root based stemmer," Proceedings of the International Arab Conference on Information Technology, 2006.

[3] S. Ghwanmeh, S. Rabab'Ah, R. Al-Shalabi and G. Kanaan, "Enhanced algorithm for extracting the root of Arabic words," Sixth International Conference on Computer Graphics, Imaging and Visualization, pp. 388-391, 2009.

[4] M. N. Al-Kabi, S. A. Kazakzeh, B. M. Abu Ata, S. A. Al-Rababah and I. M. Alsmadi, "A novel root based Arabic stemmer," Journal of King Saud University-Computer and Information Sciences, pp. 94-103, 2015.

[5] K. Taghva, R. Elkhoury and J. Coombs, "Arabic stemming without a root dictionary," In Information Technology: Coding and Computing, International Conference, IEEE, pp. 152-157, 2005.

[6] R. Alshalabi, "Pattern-based stemmer for finding Arabic roots," Information Technology Journal, pp. 38-43., 2005.

[7] R. Al-shalabi and M. Evens, "A computational morphology system for Arabic," In Proceedings of the Workshop on Computational Approaches to Semitic Languages. Association for Computational Linguistics., pp. 66-72, 1998.

[8] Q. Yaseen and I. Hmeidi, "Extracting the roots of Arabic words without removing affixes," Journal of Information Science, pp. 376-385, 2014.

[9] I. I. Hmeidi, R. F. Al-Shalabi, A. T. Al-Taani, H. Najadat and S. A. Al-Hazaimeh, "A novel approach to the extraction of roots from Arabic words using bigrams," Journal Of The American Society For Information Science And Technology, vol. 61, no. 3, pp. 583-59, 2010.

[10] E. Al-shawakfa, A. Al-Badarneh, S. Shatnawi, K. Al-Rabab'ah and B. Bani-Ismail, "A comparison study of some Arabic root finding," Journal Of The American Society For Information Science And Technology, vol. 61, no. 5, pp. 1015-1024, 2010.

[11] S. Al hakeem, G. Shakah, B. Abu Saleh and N. Thalji, "Developing an effective light stemmer for Arabic language information retrieval," International Journal of Computer and Information Technology, vol. 5, no. 1, pp. 55-59, 2016.

[12] S. Khoja and R. Garside, "Stemming Arabic text," Lancaster, UK, Computing Department, Lancaster University, 1999.

[13] T. Buckwalter, "Buckwalter Arabic morphological analyzer," 2002.

[14] M. Sawalha and E. Atwell, "Constructing and Using Broad-coverage Lexical Resource for Enhancing Morphological Analysis of Arabic," In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), pp. 282-287, 2010.

[15] A. Al_Farahidi, Kitab al-'Ayn. No publication date or place.

[16] M. Ibn Manzur, "Lisan Al-Arab." No publication date or place.

[17] A.-M. Al-Zabidi, Tag Al-'Arus Min Gawahir Al-Qamus. No publication date or place.

[18] A.-A.-Q. Al-Zamahshari, Asas Al-Balagah. No publication date or place.

[19] A. A.-F. Al-Mutrazi, Al-Mugrib Fi Tartib Al-Mu'Rib. No publication date or place.

[20] A. B. Al-Razi, Mukhtar Al-Sihah. No publication date or place.

[21] A. Al-Fayyumi, Al-Musbah Al-Munir Fi Garib Al-Sharh Al-Kabir. No publication date or place.

[22] A. A.-Q. Al-Sahib Bin 'Abbad, Al-Muhit Fi Al- Luga. No publication date or place.

[23] A. N. Al-Farabi, Al-Sihah Fi Al-Luga. No publication date or place.

[24] M. Kheder, Kalemat Al-Quraan Al-kaream, 2012.

[25] R. Sonbol, N. Ghneim and M. S. Desouki, "Arabic morphological analysis: a new approach," In Information and Communication Technologies: From Theory to Applications, 3rd International Conference, IEEE, pp. 1-6, 2008.

[26] M. Sawalha and S. Salem, "Open-source resources and standards for Arabic word structure analysis: fine grained morphological analysis of Arabic text corpora," University of Leeds, 2011.