Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

Type of the Paper (Article)

A Bi-Gram Approach for an exhaustive Arabic Triliteral roots Lexicon

Ebtihal Mustafa [1,*] and Karim Bouzoubaa[2]

[1] Collage of Computer Science and Information Technology, Sudan University of Science and technology, Khartoum, Sudan; ebtihal99@hotmail.com
[2] Mohammadia School of Engineers, Mohammed Vth University, Rabat, Morocco; bouzoubaa@emi.ac.ma
* Correspondence: ebtihal99@hotmail.com

Abstract: With the rapid development of science and technology, many new concepts and terms appear, especially in English. Other languages try to express these concepts with words from their own vocabulary. In the specific case of Arabic, there are many ways to find a counterpart for a particular new concept, such as using an existing word to denote the new concept, derivation, and blending. When these methods fail, the new concepts are simply phonetically transliterated. This has the disadvantage that most of the transliterated terms do not conform to the rules of the Arabic language and lead to a distortion of the language. Some modern linguists call for using the generation strategy to translate the new terms into Arabic by using the unused Arabic roots. Therefore, it is necessary to provide a resource that contains all Arabic roots with a categorization of what is used, what is available for use, and what is rejected according to the phonetic system. This work provides a comprehensive lexicon that contains all possible Arabic triliteral roots, determines the status of each root in terms of usage and acceptability, and provides a mechanism for giving preference to roots when there is more than one root that indicates the desired meaning.

Keywords: Arabic language; Arabic roots; lexicons, phonetic system, bigram frequencies, roots weight, Artificial Intelligence, NLP, Arabic NLP.

1. Introduction

The Arabic language is one of the oldest languages that originated in the Arabian Peninsula in pre-Islamic times. It belongs to the Semitic family along with Amharic, Aramaic and Hebrew. It is the most widely spoken and studied language in this family (Al-Huri 2015) and also the religious language of all Muslims.

Arabic consists of 28 letters of the alphabet and is a highly derivational language. The vocabulary of Arabic words is essentially derived from roots. These roots may consist of three, four, or five letters, such as ك ت ب (ktb), د ح ر ج (dHrj), and س ف ر ج ل (sfrjl) (Al-kabeerm, Hasboallah, and Al-shazli 1981). Derivatives of these roots are formed by attaching affixes to the roots according to specific patterns. The patterns are used as standard frames for Arabic lexical words. For example, applying the pattern فاعِل (fAEil) on the triliteral root كتب (ktb) results in the lexical form كاتِب (kAtib/ writer).

However, not all combinations of the 28 letters are used as roots, either because they are difficult to pronounce like ع ح ه (hHE) or because there was no need to use them like ع ج ل (ljE) (Kishli 1996). The difficulty of pronunciation is a matter for phoneticians who study the letter combinations within roots. Arabic phoneticians have focused primarily on triliteral roots, since quadriliteral and quinqueliteral largely share the properties of triliteral roots. Depending on the criteria of easy pronunciation, they have divided the degree of acceptability of the phonetic combination into three types: suitable, less suitable, and unsuitable (Alfozan 1989).

With the suitable type, there is no difficulty in pronunciation because the sounds are articulated far apart. An example of this type is the root م ل أ (alm). The less suitable type

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

contains two identical letters such as م ك ك (makk) and س ب ب (sabb). The unsuitable type is the one that contains sounds that are difficult to combine because they are articulated very closely, especially those that are articulated in the throat, such as ع ح ه (hHE) (Hindawi 1993).

Unfortunately, the current situation regarding Arabic roots and corresponding words is not keeping pace with the continuous scientific development. In fact, many new terms are emerging with the development in science and technology and all fields of life. Some studies have estimated that more than (50%) of the vocabulary of developed countries are scientific terms (Dwaidri 2010). Consequently, many countries are trying to follow the scientific trends and are making efforts to expand their languages to accompany this development.

Concerning Arabic, there are several strategies for dealing with new terms. These strategies are: (1) modifying the original concept of an existing word to adopt the new concept, such as سيارة (syArp; car), since in ancient times the Arabic word had the meaning of a group of walking people or a convoy (Al-kabeerm, Hasboallah, and Al-shazli 1981), while today it is more commonly known as a car, (2) Arabizing foreign words according to the Arabic forms (al-taʿrib; Arabization) such as تلفاز (tilfaz; television), (3) blending or merging of two words into one (al-naht; blending) such as برمائي (brmaai; amphibious), and finally (4) deriving new expressions from original Arabic roots (al-iŠtiqāq; derivation) such as حاسوب (HAswub; computer), which is a new word derived from the root (ح س ب/ Hsb/ compute) (Brakhw and Milad 2019).

Modifying the original meaning of the word to fit the new concept is one of the most effective methods of creating new terms and the resulting term is easy to understand, but sometimes it is not possible to have an old Arabic word that is suitable for the new intended meaning, so the new term must be created using one of the other methods.

Arabizing may produce words that do not conform to the phonetic system of Arabic. For example, the term "hydroxy" is translated as "هيدروكسي" whose pattern "فيعلوللي" is not Arabic. Moreover, in Arabization, it is not possible to maintain the relationship between the Arabic root and the Arabized term (Al-Shbiel 2017). For example, using the Arabic term محرك (muHarik) as an equivalent for the English term motor, associated with the Arabic root ح ر ك (Hrk/ move). The Arabicized term موتور (mwutwur/motor), on the other hand, is not associated with any Arabic root.

Blending plays an effective role in handling affixations and abbreviations of long Arabic terms such as (لافقاري; invertebrate) and (كهروميغناطيسي; electromagnetic). However, there are restrictions on the use of blending and it may only be used for scientific necessity (Elmgrab 2011). These restrictions are due to the fact that in blending, there are no rules that must be followed during the process, while Arabic has specific rules and patterns that cannot be eliminated (Ali Al-foadi 2018).

derivation is the best choice as suggested by many authors/works (Ali Al-foadi 2018). Indeed, as mentioned above, modifying modifying the old word meaning to fit the new one does not always work, and the terms created by Arabization and blending may be incompatible with Arabic, increasing of such terms leads to distortion of the language (Al-Salih 1968). Therefore, some lexicographers suggest using new roots for new terms by deriving the corresponding Arabic words from these new roots.

It is worth noting that the methods for generating terms were proposed by linguists and not handled by natural language processing (NLP) researchers. As described in the Related Work section, most of the research in the NLP field concerned either collecting statistics on the used roots or studying the phonetic system of the Arabic language. However, the results of these efforts were not exploited in generating new terms.

In order to help lexicographers propose new Arabic scientific terms using the generation strategy, and since there is still space from all roots' combinations, this study aims to develop an algorithm that generates all possible triliteral roots, determines whether they are used or not, are phonetically accepted or not, and to what extent they are compatible with the phonetic system of the Arabic language. These roots can then be com-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

bined with patterns to generate new lexical forms that can be evaluated by lexicographers.

The rest of the paper is as follows. Section 2 reviews previous works. Section 3 describes the methodology and section 4 shows the results. The paper concludes in section 5.

2. Related Work

Arab scholars have been interested in lexicography since ancient times and have excelled in this field both in variety and perfection. They have used various methods to collect and arrange vocabulary in lexicons (Omer 1995). Among the most famous Arabic lexicons are Al-Sahah (Attar 1987), Lisan Al-Arab (Al-kabeerm, Hasboallah, and Al-shazli 1981), Taj Al-Arous (Shiri 1994), Al-Wassit (Anees et al. 2004), and Al-Moassir (Omer 2008). In most of these lexicons, the vocabulary is divided into groups; each group belongs to the root from which it is derived. For example, the words مدرسة (mdrsp/school), دراسة (drAsp/study), and دارس (dArs/student) belong to the root د ر س (drs). The size of the vocabulary varies from one lexicon to another due to the differences in time and method of its collection; in each time period, some terms appear and become popular, while the use of others decreases or disappears. Table 1 shows statistics about roots of the mentioned Arabic lexicons.

Table 1. lexicons statistics

	Triliteral	Quadriliteral	Quinqueliteral	Total
Al-Sahah	4,814 – 86% of total	766	38	5,618
Lisan Al-Erab	6,538 – 71%	2,548	187	9,273
Taj Al-Arous	7,597 – 63%	4,081	300	11,978
Al-Wassit	5,155 – 78%	1,332	153	6,640
Al-Moassir	3,292 – 67%	1,092	535	4,919

Table 1 shows that the most comprehensive lexicon is Taj Al-Arous, and the triliteral roots are the most used in the language. However, if we compare the used roots in these lexicons with the total number of roots that can be formulated from the twenty-eight letters of the Arabic language, it becomes clear that there is a large gap between them, as shown in Table 2.

Table 2. Letters Combinations Statistics

	Possible roots	Used roots in taj al arous	Percentage
Triliteral	21,952	7,597	34.6 %
Quadriliteral	548,800	4,081	0.74 %
Quinqueliteral	9,765,625	300	0.003 %

The first scholar to notice the gap between the possible roots and the used root was Al-Khalil, who called this phenomenon "Al Muhmal" (i.e., the unused) and explained that it is caused by difficulties in pronunciation (Kishli 1996). Many linguists after Al-Khalil also studied this phenomenon to discover the reasons for the unused combinations. Ibn Duraid (Balabaki 1987) added the "disharmony of letters" as another reason. Ibn Jinni (Hindawi 1993) justified the unused combinations by arguing that there is no need for such terms or that there is a lack of unison between the sounds that make up the root and the intended meaning.

On the other hand, NLP researchers also studied and analyzed lexicons to know how Arabic words are formulated in order to use them in developing and expanding the language or to ensure the accuracy of the results obtained by ancient scholars. The first use of computers in Arabic linguistics was in the 1970s when (Musa 1978) conducted a statistical study on the roots of the Al-Sahah lexicon to investigate some linguistic phenomena. Al-fozan (Alfozan 1989) also devoted a research to study, enumerate and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

summarize the impossible combinations from ancient books. He collected more than 80 phonetic rules and corrected some rules addressed by Al-khalil ibn Ahmed, such as the combination of the letters " أ/ >" and "هـ / h" which are combined in the root "أهل/ > hl".

(Alm and Al-Faham 1983) studied the combination of letters using the bigram frequencies of Arabic roots. They wanted to verify the accuracy and completeness of the results obtained by Al-Khalil with respect to letters that could not be combined in any Arabic root and resulted in many combinations not being used. By performing their experiment on five lexicons, they proved the strength of Al-Khalil's results.

(Hegazi 2016) also shed light on the gap between the possible roots and the roots by creating a lexicon that includes all possible triliteral roots. In this lexicon, Arabic roots are generated by applying permutations to the Arabic letters. Then he applied the Arabic patterns to the roots to obtain the words or vocabulary. The drawback to Hegazi's study is that he did not consider the combinations that were not used. He excluded only the roots that consist of three redundant letters. He applied Arabic patterns to all 21924 roots.

As we will show, NLP studies were conducted not only in theoretical terms, but also in practice. For example, (Abdoalrasool 2010) exploited the unused combinations in the context of optical character recognition to improve the quality of the output depending on the Arabic language features without using spell checking or morphological analysis. Also, Mai in (Abusair 2012) and Al-Radaideh in (Al-Radaideh and Masri 2011) used Arabic bigrams in prediction to improve writing Arabic SMS messages on 12 keys cell phones.

As we have already mentioned, languages are constantly evolving and expanding with the development of life. The Arabic language, as one of the most widespread languages, must keep pace with this development and evolve in its own way without blurring its characteristics. In this regard, there are many studies deal with new terms. For example, (Elmgrab 2016) tried to find a suitable technique for creating new terms in Arabic, and (Hassan 2017) proposed an approach to automatically translate the new terms into Arabic. In both studies, it was found that the best strategy for introducing new terms in Arabic is derivation.

There are many researches conducted either by linguists or NLP researchers on Arabic lexicons. Most of them served the purpose of studying linguistic phenomena that characterize the Arabic language, while others used these phenomena for useful applications. However, the obtained results were not used to extend the language especially through NLP tools. The work in this area was limited to linguists, although the use of NLP tools will greatly help.

In this context, Rajaa (Dwaidri 2010) points out in her book the necessity to create a bank for all Arabic roots in order to unify the Arabic lexicon to be used in expanding the language. This bank must include existing Arabic roots from well-known lexicons such as the Al-Sahah, Lisan Al-Arab, and Taj Al-Arous, as well as the unused and phonetically rejected roots.

In this work, an attempt is made to use these phenomena to create a lexicon that will help the lexicographers who use the generation strategy to propose new Arabic terms to develop and expand the language.

3. Methodology

This study presents an approach to generating all Arabic triliteral roots. For each generated root, we determine whether it is used in Arabic or not. For the unused roots, we explain whether they are accepted or not according to the Arabic phonetic system. Then, we assign a weight to each root indicating how much this root is compatible with the Arabic phonetic system. To do this, we proceed in several steps, as shown in Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

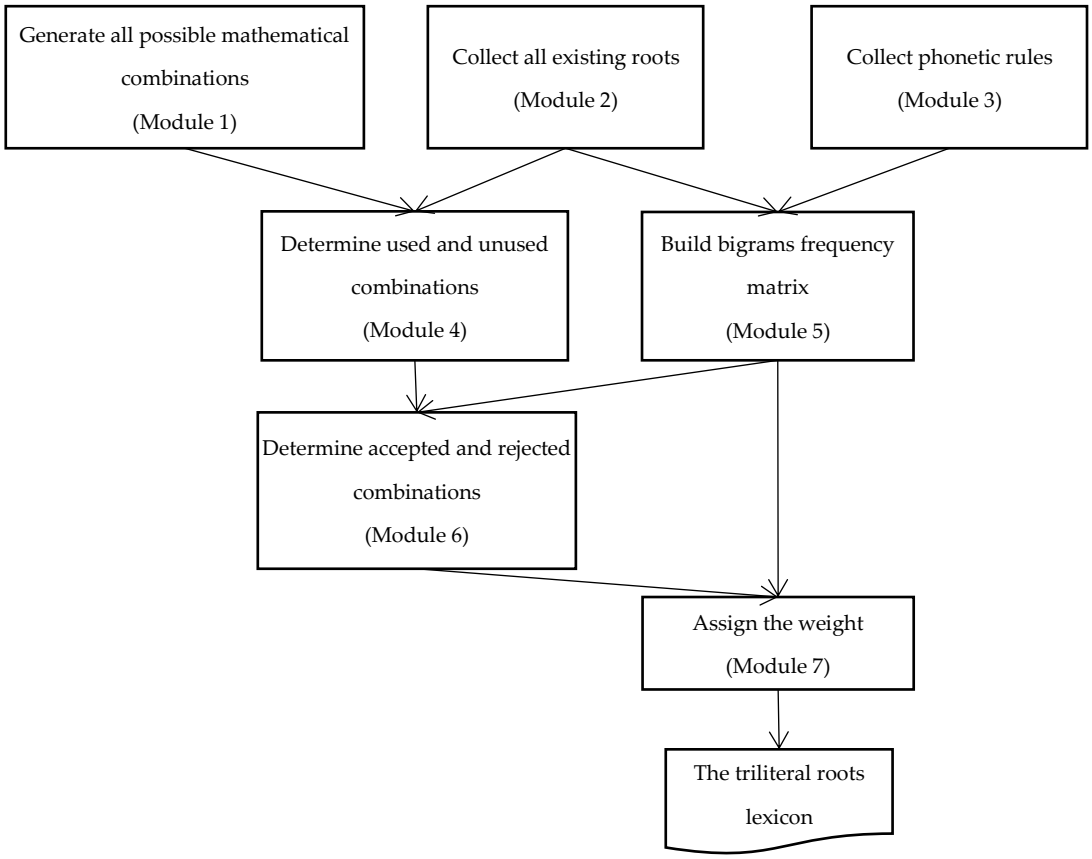

Figure 1. Proposed Approach

The first three modules are independent and can be run in parallel. Module 1 is used to generate all combinations consisting of three of the 28 Arabic letters. Module 2 collects existing roots from the lexicons. Then, the output of Module 1 and Module 2 are passed to Module 4 in order to mark each generated root from Module 1 as used or unused according to the output of Module 2.

Module 3 and Module 5 collect the letters that cannot be combined in a root. Most of these impossible letter combinations are addressed in ancient Arabic books, and the unaddressed ones are extracted from the existing roots.

According to the output of Module 5, Module 6 marks each generated root as accepted or rejected. Module 7 assigns a weight to each root to indicate how much this root is compatible with the Arabic phonetic system. Finally, we obtain a lexicon that contains all mathematically possible triliteral roots, which are assigned specific labels such as the acceptance and usage of the root in the language.

Figure 2 shows a simple example of the output of each module from Figure 1. All modules are explained in detail in the next subsections.

3.1. Generating All Roots

The proposed generation algorithm is based on mathematical permutations and combinations where all possible triliteral combinations of the twenty-eight Arabic letters are generated reaching a total number of 21,952 combinations (28 x 28 x 28).

The first generated root in Module 1 is "ااا/ >>> ", followed by "ااب/ >> b" until it ends with the root "يييي/ yyy". Some of the generated roots are already used in Arabic, while others are not. To determine whether a root is used or not, we consider the five mentioned lexicons as explained below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

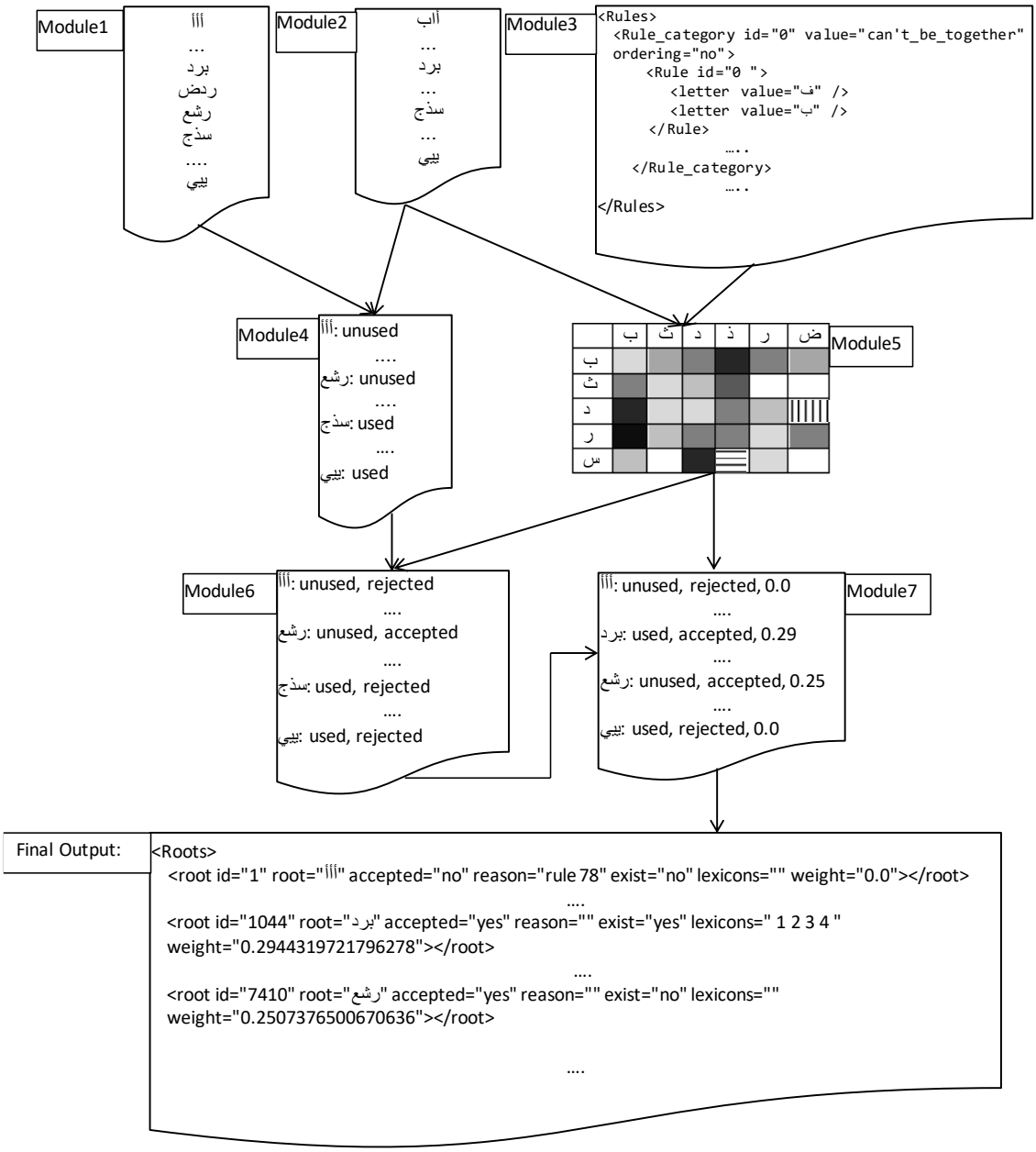

Figure 2. Modules Output Samples

3.2. Collecting the existing roots

The triliteral roots are collected from five selected lexicons as already shown in Table 1 assuming that they ensure completeness. After merging their roots and removing redundancy, we obtain 8,426 distinct ones.

The existing roots are collected (in Module 2) not only to distinguish between used and unused ones (from Module 1), but also to obtain information about the phonetic system from the practiced language and how letters are combined to formulate the roots.

Figure 2 shows that the first existing triliteral root following the alphabetical order, is "أاب/ <Ab" while the last one is "ييي/ yyy".

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

3.3. Collecting phonetic rules

As mentioned earlier, some letters cannot be combined in a root because of the difficulty of their pronunciation or their incompatibility with each other such as the letters "س/s" and "ث/v". The roots that contain such an impossible combination are phonetically not accepted and must be excluded; this means there are phonetic rules that control the acceptance of the root in the language. These unused combinations were used as phonetic rules to recognize Arabicized roots.

As explained in the previous section, there are a number of modern linguists interested in the collection of phonetic rules (Alfozan 1989, Alm and Al-Faham 1983). Our effort in this regard is to organize the phonetic rules and put them into a standardized digital format that is accessible to everyone and easy to use. We have put all the addressed phonetic rules in an xml file. Each rule has an ID, a category, and letters that cannot be combined according to the rule. Figure 3 shows an example of the phonetic rules file.

```
<Rules>
    <Rule_category id="1" value="can't_be_together" ordering="no">
        <Rule id="1 ">
          <letter value="ف" />
          <letter value="ب" />
      </Rule>
            …..
    </Rule_category>
    <Rule_category id="2" value="can't_be_followed_by" ordering="yes">
        <Rule id="32 ">
          <letter value="س" order="1"/>
          <letter value="ش" order="2"/>
      </Rule>
            …..
    </Rule_category>
    <Rules_category id="3" value="composed_of_identical_letters">
      <Rule id="50" lett1="ا" lett2="ا" lett3="ا"></Rule>
            …..
    </Rules_category>
    <Rules_category id="4" value="start_with_identical_letters">
      <Rule id="78" lett1="ا" lett2="ا" ></Rule>
            …..
    </Rules_category>
</Rules>
```

Figure 3. Phonetic rules xml file example

As can be seen in Figure 3, there are four categories of rules; the last two categories contain phonetic rules that apply to all letters, namely that the root must not consist of three identical letters "composed_of_identical_letters" and must not start with two repeating letters "start_with_identical_letters" such as "ففف\ fff" and "ففر/ ffr" respectively.

The first two categories, "can't_be_together" and "can't_be_followed_by," on the other hand, contain rules that prevent the co-occurrence of some letters in a root. For

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

example, the letters "ف/f" and "ب/b" cannot be combined in a root, regardless of their order. So, this rule belongs to "can't_be_together" category, where it does not matter which of the two letters precedes the other.

The letter "د/d" cannot be followed by the letter "ت/t" in any root, whereas the letter "ت/t" can be followed by the letter "د/d" as in "وتد/wtd". So, this rule belongs to "can't_be_followed_by" category, where the letters can be combined in a root only in a certain order.

Nevertheless, not all phonetic rules are addressed in the ancient books due to the lack of capabilities in that time. Therefore, the unaddressed rules are extracted by analyzing the combinations in existing roots using a bigram frequency matrix. This is explained in more detail in the next section.

3.4. Building bigrams frequency matrix

In the context of natural language processing, a bigram is a sequence of two adjacent elements from a string of tokens, which are usually letters, syllables, or words. The frequency distribution of each bigram in a string is used in many applications such as computational linguistics and speech recognition for statistical analysis of text.

To obtain the bigram frequencies from Arabic lexicons, a 28x28 matrix is created. Each row and column represents an Arabic letter. The cell where the rows and columns intersect indicates how often these two letters occur in all entries of the lexicon.

To fill the matrix, from the five selected lexicons each root split into three bigrams. For example, the root كتب is split into تب ,كت and كب and the cell corresponding to each bigram is incremented by one. The corresponding cell for the bigram تب is the cell located at the intersection of the row representing the letter ت and the column representing the letter ب.

For a more detailed representation, we obtain three matrices. The first matrix represents the first bigram (the bigram representing the letters in the first and second positions, كت in the previous example), the second matrix represents the second bigram (تب in the previous example), while the third matrix represents the first and third bigram (كب in the previous example). Moreover, we can combine the three matrices into one matrix to get a global view of the frequency of bigrams. The bigram frequency matrix is statistically known as the correlation matrix.

The bigram frequency matrix can be represented in the form of a heatmap. It is a graphical representation of data where values are represented by colors and/or textures. The heatmap makes it easier to visualize the data and understand it at a glance. Figure 4 shows the correlation matrix between Arabic bigrams extracted from five lexicons and visualized using the heatmap.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

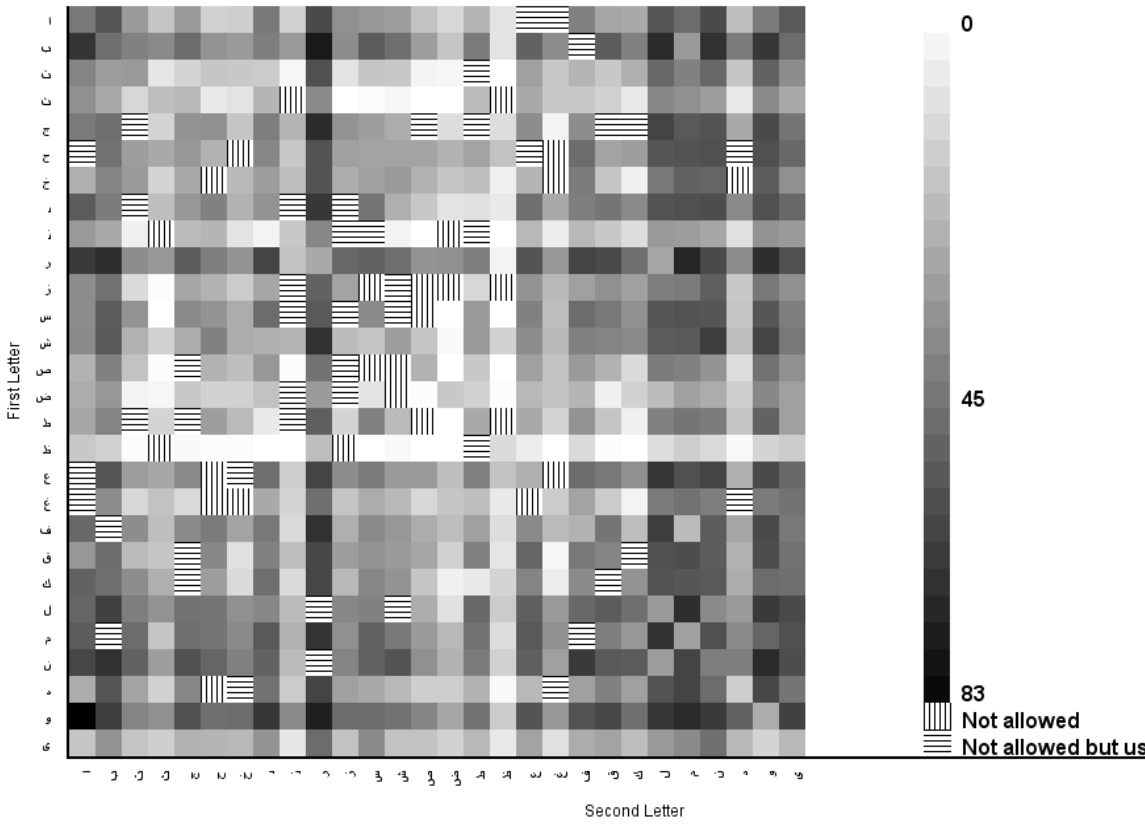

Figure 4. Arabic Roots Bigram Frequencies

The darkest cell means that this bigram is more frequent, and the frequency decreases when the cell is lighter. A white cell means that the corresponding bigram does not occur in any root. For example, the frequency of the bigram طب is less than the frequency of the bigram رب, the frequency of the bigram ذهـ is less than the two bigrams that mentioned before, while the frequency of the bigram ظش is zero (shown in white), which means that there is no existing root containing ش and ظ.

As explained earlier, in order to obtain all phonetic rules, we cannot rely only on the addressed rules, since they are not complete. We also cannot rely only on the root analysis result, since some Arabicized roots contain impossible letter combinations, such as "سذج/s*j", and such roots may affect the analysis result. Therefore, the root analysis process must include information about the addressed phonetic rules to avoid the effects of such exceptions. Therefore, the addressed rules appear in the bigrams frequency matrix as vertical and horizontal stripes.

The vertical stripes indicate that the corresponding bigram is not allowed by the addressed phonetic rules of Arabic, such as ثذ. However, some Arabicized roots may contain non-allowed bigrams and such cases are indicated by horizontal stripes. For example, although there is a rule prohibiting the combination of س and ذ in a root, we find the bigram سذ extracted from the Arabized root سذج found in the four of the five selected lexicons.

As mentioned above, there are many bigrams in Arabic that cannot occur together in one root. Many of them are addressed in Arabic books and are denoted in vertical stripes in the heatmap. However, the heatmap helps identify unaddressed bigrams because they are shown in white color. The addressed phonetic rules are about 84 rules, while the rules extracted from the matrix are 107 rules; this means there are more than 20 rules that are not addressed. We create an xml file with the Arabic phonetic rules, whether they are addressed or obtained by analyzing the used roots[1].

[1] The file is available at http://arabic.emi.ac.ma/alelm/?q=Resources

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

To find out whether the generated root of Module 4 is phonologically acceptable or not, we divide the root into bigrams and then compare these bigrams with the bigram frequency matrix. If one of the root bigrams corresponds to the white or striped cell, then that root is phonetically unacceptable, otherwise it is phonetically acceptable. For example, in Figure 2, the bigrams "عح" ,"ثض", and "مف" are phonetically unacceptable because they correspond respectively to white, vertically striped, and horizontally striped cells.

As previously mentioned, phoneticians divide the degree of acceptability of root sounds into three types: suitable, less suitable, and unsuitable. This means not all phonetically acceptable roots have the same degree, but some are preferred over the others depending on the letters that compose them. Next step is to assign a weight to each root expressing this degree in numbers.

3.5. Assigning the weight

As previously explained, ease of pronunciation has been expressed by linguists in rules representing the possibility of coexistence (or non-coexistence) of two letters in a root. Therefore, the idea is to calculate the weight of a root by calculating the weight of each of its three bi-grams. To do so, we first assign a weight to each bigram individually and then combine these weighs to calculate the global weight of the root. To assign a weight to a bigram, we use probability theory (Sherlock and Ormell 1970). The weight of the bigram is calculated as follows:

$$w_{(xy)} = \frac{freq_{(xy)}}{freq_{(bi)}} \quad (1)$$

Where:

$w_{(xy)}$: weight of the bigram (xy)

$freq_{(xy)}$: frequency of the bigram (xy)

$freq_{(bigrams)}$: frequencies of all bigrams

The frequency of the bigram is obtained from the corresponding cell in the bigram frequency matrix, while the frequency of all bigrams is obtained by summing the frequencies from the corresponding matrix.

After assigning a weight to each bigram, the next step is to aggregate these weights into a value that is assigned to the root. The aggregation formula consists of multiplying the weights of the bigrams as in the following equation.

$$w_{(root)} = (w_{12} * w_{23} * w_{13}) \quad (2)$$

Where:

$w_{(root)}$: weight of the root

w_{12}: weight of the first and second letters bigram

w_{23}: weight of the second and third letters of bigram

w_{13}: weight of the first and third letters bigram

Multiplication ensures that the value of the total weight of the root is high only if the values of all bigrams are high, and if one bigram is un- accepted, the value of the root weight is zero.

According to the proposed weighting scheme, the unused roots "حتس" and "رشع" have a weighting value of 0.01 and 0.07, respectively, while the used roots "رأي" and "شرب" have a weighting value of 0.54 and 0.37, respectively, and the roots that violate any of the phonetic rules have a weighting value of zero, regardless of whether they are used in the Arabic language or not.

4. Results

Through this work we want to build a lexicon containing all triliteral combinations, and determining which ones are phonetically rejected which ones are used, and which ones are available to be used by linguists to extend the language. To achieve our main goal, we have gone through several stages; some of them had intermediate results such as Arabic phonetic rules file. These results are available to researchers in the field as ex-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

plained earlier. The main result is a lexicon of triliteral roots, as shown in Figure 5, where each root has several attributes.

```
<Roots>
    <root id="1" root="ءءء" accepted="no" reason="rule 1" exist="no" lexicons="" weight="0.0"></root>
                        ……
    <root id="29" root="أبأ" accepted="yes" reason="" exist="yes" lexicons=" 1   2   4 " weight="0.68"></root>
    <root id="30" root="أبب" accepted="yes" reason="" exist="yes" lexicons=" 1   2   3   4   5 "
weight="0.25"></root>
                        …..
    <root id="7410" root="رشع" accepted="yes" reason="" exist="no" lexicons="" weight="0.07"></root>


                        …..
    <root id="8853" root="سذج" accepted="no" reason="rule 84" exist="yes" lexicons=" 1   2   4   5 "
weight="0.0"></root>
                        …..
    <root id="21952" root="ييي" accepted="no" reason="rule 28" exist="yes" lexicons=" 1 " weight="0.0"></root>
</Roots>

<Rules>
    <Rules_Category cat_id="1" value="composed of identical letters">
        <Rule id="0" letter1="ا" letter2="ا" letter3="ا"></Rule>
      ….
    </Rules_Category>
      ….
</Rules>

<Lexicons>
```

Figure 5. Structure and contents sample of the proposed lexicon

The first attribute "id" is the root's identification number, which has a value between 1 and 21952, based on the root's alphabetical order. The "root" attribute is a three-letter combination of Arabic letters. The "accepted" attribute determines whether the root is acceptable or not according to the phonetic system of the Arabic language. If the root is phonetically rejected, the reason for the rejection is explained in the "reason" attribute by specifying the ID of the phonetic rule that the root violated.

The "exists" attribute determines whether the root is used in the language and is present in the Arabic lexicons or not. If it is used, the lexicons attribute contains the IDs of the lexicons that contain the root. If the root is not used, the value of the lexicons attribute is empty. The last attribute is the "weight", whose value determines the compatibility of the root with the phonetic system of the Arabic language. If the root violates one or more phonetic rules, the value of the weight attribute is zero, even if the root is used.

For example, the first root "ءءء" has id 1 and is not accepted according to rule 1, which states that the root must not consist of three repeating letters. This root is not used in Arabic, so the value of the "exists" attribute is equal to no and the value of the "lexicons" attribute is empty. Since the root violates a phonetic rule, the value of the "weigh"t attribute is zero.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 April 2022 doi:10.20944/preprints202204.0128.v1

As previously mentioned, not all triliteral combinations are used. Some of them are not subject to the Arabic phonetic system, while the others are phonetically accepted. Some linguists have advocated using these roots to expand the language rather than borrowing a large number of terms that could blur the language.

Applying permutation to the twenty-eight letters of the Arabic alphabet yields 21,952 three-letter combinations that can be divided into two main categories: phonetically accepted and phonetically rejected. Each of these categories is in turn divided into used and unused. This results in four categories: phonetically accepted used category, phonetically accepted unused category, phonetically rejected used category, and phonetically rejected unused category. Table 3 provides statistics for each of these categories.

Table 3. Three Letters Combinations Statistics

All possible combinations 21,952			
Phonetically accepted combinations 13,410		Phonetically rejected combinations 8,542	
Unused 5,383	Used 8,027	Used 399	Unused 8,143
	8,426		

The phonetically accepted used category (8,027) forms the vast majority of the current language; the phonetically rejected used category includes the exceptions in the current language (399) such as Arabicized roots. Thus, the current Arabic language uses 8,426 (8,027+399) forms. In turn, the phonetically rejected unused roots category (8,143) contains the roots that do not follow the Arabic phonetic system and are not used, and the phonetically accepted unused category (5,383) can be used to expand the language.

This last number (more than 5,300) shows that there is a wide range of roots that are accepted and not used and can be used to extend the language. If we compare the number of words in the lexicon with the number of roots, there are on average 13 words derived from a root. This means that unused roots can produce as many as 70,000 new words. Words are generated from the accepted - unused - roots by applying Arabic patterns. For example, some of the words that can be derived from the root "رشع/r$E" are "راشع/rA$iE", "مرشوع/mr$uwE", "مرشعة/mir$Ep", "مرشاع/mir$AE" and "رشعة/r$Ep".

5. Conclusions

In this paper, we have attempted to provide researchers with a comprehensive triliteral root lexicon containing information on what is used, what may not be used, and what can be used to extend the language. We relied on mathematical combination and permutation theory to generate the roots to ensure that all roots are processed. Then we merged five Arabic lexicons to know which roots are actually used. To determine the acceptability of each root, we used a bigram frequency approach based on the merged lexicon to create a corresponding heatmap matrix. In addition to the linguistic addressed phonetic rules, this matrix is used to (i) extract other phonetic rules on the one hand, and (ii) calculate the weight of the roots on the other hand, which indicates the compatibility of the root with the Arabic phonetic system. The results show that there is a large space of available combinations that can be used by linguists to extend the language. Future research is needed to determine how this space can be used by researchers to extend the language and how to assign meaning to each root.

 doi:10.20944/preprints202204.0128.v1

References

Abdoalrasool, Amro Jumaa. 2010. "Tatweer Alta'rof Alali Ala Alhoroof Alarabiea Min Khilal Aliea Loghawiea." In *International Computing Conference in Arabic*, edited by Yasmine Hammamet, Moncef Charfi, and Hani Ammar. Tunisia: Phillips Publishing, Ohillipsburg, NJ. http://www.phillips-publishing.com/.

Abusair, Mai I. 2012. "Improving Arabic Text Entry Methods Using Word Bigrams Prediction And Keys Reassignment." In *International Conference on Intelligent Computational Systems*. Dubai.

Al-Radaideh, Qasem A, and Kamal H Masri. 2011. "Improving Mobile Multi-Tap Text Entry for Arabic Language." *Computer Standards \& Interfaces* 33 (1): 108–13.

Alfozan, Abdulrahman Ibrahim. 1989. "Assimilation in Classical Arabic: A Phonological Study." University of Glasgow.

Ali Al-foadi, Raheem. 2018. "Derivation as the Main Way of Adapting New Terms to Arabic." *Modern Journal of Language Teaching Methods (MJLTM)* 8 (3): 194–99.

Alm, yahya meer, and Shakir Mohammed Al-Faham. 1983. "Derasa Ehsaiea Lidwaran Alhoroof Fi Aljozoor Al-Arabiea." Damascus university.

Balabaki, Ramzi Monir. 1987. *Jamhrat Al-Logha Li Ibn Duraid*. 1st ed. Dar al-ilm.

Dwaidri, Rajaa Waheed. 2010. *Al-Mostalah Al-Elmi Fi Al-Logha Al-Arabiea, Omqaho Al-Turathi Wa Boadho Al-Moassir*. 1st ed. Damascus: Dar al-fikr.

Elmgrab, Ramadan Ahmed. 2016. "The Creation of Terminology in Arabic." *American International Journal of Contemporary Research* 6 (2): 75–85.

Hassan, Sameh Saad. 2017. "Translating Technical Terms into Arabic: Microsoft Terminology Collection (English-Arabic) as an Example." *Translation \& Interpreting, The* 9 (2): 67–86.

Hegazi, Mohamed Osman. 2016. "An Approach for Arabic Root Generating and Lexicon Development." *Int. J. Comp. Sci. Netw. Sec.(IJCSNS)* 16 (1): 9.

Hindawi, Hassan. 1993. *Sir Sinaat Al-Erab Li Ibn Jinni*. Damascus: Dar al-qalam.

Kishli, Ḥikmat. 1996. *Kitab Alain Lil-Khalil Ibn Ahmed Al-Farahidi*. Bayrut: Dar al-Kutub al-ilmiyah.

Musa, Ali Hilmi. 1978. *Dirasa Ihsaeia Lijzoor Muajm Al-Sahah Bistikhdam Al-Computer*. 1st ed. Ciro: Al-haiaa Al-masriea Al-ama lilkitab.

Sherlock, Alan, and C. P. Ormell. 1970. *An Introduction to Probability and Statistics*. *The Mathematical Gazette*. Vol. 54. https://doi.org/10.2307/3613205.