Journal of Theoretical and Applied Information Technology
15th March 2023. Vol.101. No 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# USING STRING SIMILARITY ALGORITHMS TO FIND ENGLISH WORDS POTENTIALLY ORIGINATING FROM ARABIC

**MAJED ABUSAFIYA**[1]

[1]Associate Professor, Al-Ahliyya Amman University, Department of Software Engineering, Jordan

E-mail: [1]mabusafeyeh@ammanu.edu.jo, majedabusafiya@gmail.com

**ABSTRACT**

One interesting linguistic issue is the study of English words originating from Arabic. This is based on identifying pairs of Arabic and English words that show obvious similarity in pronunciation and meaning. To the best of the knowledge of the author, no computational solution was proposed to support this subject. In this paper, similarity algorithms are used to measure the similarity between Arabic roots and English words. The proposed solution is implemented and the resulting similarity data was studied. One contribution of this work is finding many pairs of Arabic and English words that did not exist in the widest collective reference dictionary that is known in this context. Another contribution is showing how the string similarity algorithms may be utilized to this purely linguistic issue.
**Keywords:** *Algorithms, String Similarity, Computational Linguistics, Arabic, English*

## 1. INTRODUCTION

Many languages borrowed words from Arabic. For example, more than one hundred and sixty words are explicitly stated to be from Arabic origin in Oxford English Dictionary [1]. A number of studies focused on this issue. For example, the author of the *Paradise Dictionary* [2] listed more than 10,000 English words claiming that they originate from Arabic. This dictionary is considered the widest and the most collective reference in this context. Author of [3] stated that there are hundreds of loan words in English that are from Arabic. He stated that the claim in [2] is exaggerated since little evidence was presented to support this claim. In addition to that, there was little attention given to this work from English linguistics researchers. The author of [4] admitted that he cannot agree or disagree about the claim in [2]. However, he suggested more detailed studies to verify it. A historical dictionary showing the contributions of Arabic to English may be found in [5]. Claiming that English is the one that is borrowing from Arabic and not vice versa is supported by many facts: (1) Arabic is much older and richer in roots than English. For example, the first Arabic dictionary - *Al-ayn* (العين) - was written in the seventh century by *Al-khalil Al-faraheedi*. It contains more than ten thousand roots. On the other hand, the oldest English dictionary - *A Table Alphabetical* - was written in the beginning of the sixteenth century by *Robert Cawdrey's*. It contains about 3000 words only. (2) Isolation of Arabs from other nations makes it less likely for Arabs to borrow words from other nations. (3) Arabic preserved its original shape for centuries while other languages shows continuous change.

Claiming that a given English word may have originated from Arabic may be supported by identifying a corresponding Arabic word with very close pronunciation and meaning. The focus of this paper is to present a computational solution that may help in identifying these words using string similarity algorithms. A literature review for studies that used computation to help in finding these pairs was conducted. No such work was found by the author. For a string similarity algorithm to work in this context, the comparison should happen between strings of the same alphabet. For this reason, a transliteration of words from one language to another is required. To get more accurate similarity measures, a dual transliteration is carried out: for English words to Arabic and for Arabic words to English. Similarity algorithms are then applied on both pairs and the higher similarity value is considered for that pair. To cope with shortcomings of using a single similarity algorithm, seven different string similarity algorithms are used for measuring similarity. The computed similarity values for the seven algorithms for a pair of Arabic and English words are stored in a sorted vector. The generated similarity data was studied for the Arabic

roots for the first five Arabic alphabet letters from the selected Arabic dictionary. Almost half the pairs that were found in this study did not exist in *Paradise* dictionary. This proves the value of the proposed approach in finding such pairs.

This paper is structured as follows: in section 2, the main algorithm is presented. Section 3 introduces the string similarity algorithms that were used. Section 4 presents the process that was followed by the author to analyze the similarity data that was generated by the implemented system along with the discovered results. The paper ends with a conclusion and a list of references.

## 2. THE MAIN ALGORITHM

The main algorithm to measure the similarity between Arabic and English words is shown in Fig. 1. Arabic words that are used in this comparison are the roots of Arabic that are registered in Maqāyīs Al-Lugha dictionary (معجم مقاييس اللغة) [5] - which contains 5,306 roots. Restricting Arabic words to the set of Arabic roots is justified because it is very hard to run this algorithm on all Arabic words. Arabic words are so numerous to be gathered or enumerated. Another advantage of using the Arabic roots is that they contain only the basic letters that are associated with a particular meaning. This is due this to specific feature of Arabic where its individual letters are associated with meanings.

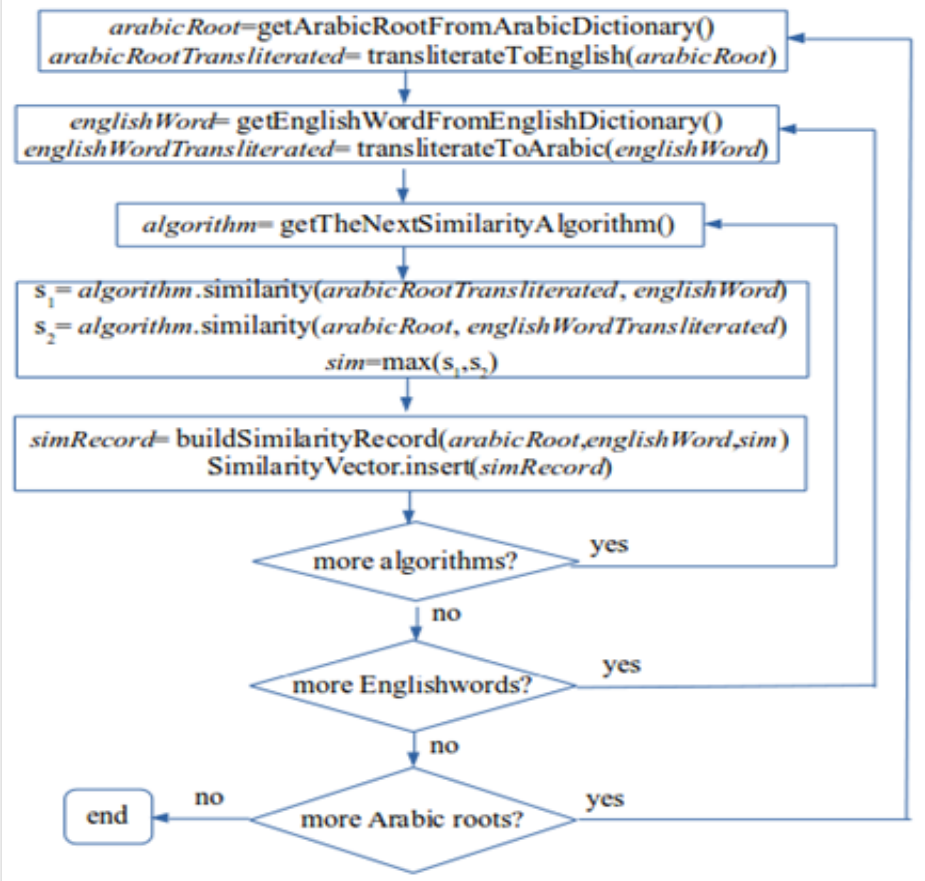


*Figure 1: The Main Algorithm*

English words that are considered in this study are those that are indexed in Oxford dictionary. The soft copy that was used in this research contains 36,639 words. The number of words in Oxford dictionary is much larger than that of the selected Arabic dictionary because it is not restricted to the roots. The algorithm takes Arabic roots one by one and transliterates them to English Alphabet. For example, the Arabic root (كحل) is transliterated to (*khl)*. Then, words in Oxford dictionary will be taken one by one and then transliterated to Arabic alphabet. For example the

English word (*kohl*) will be transliterated to (كوهل). Using each of the seven selected similarity algorithms, the similarity between the current Arabic root and the Arabic transliteration of every English word will be measured. Also these algorithms will be run to measure the similarity between the English transliteration of the current Arabic root and every English word. Therefore, for a pair of given Arabic root and English word, two similarity values will be calculated for each of the seven string similarity algorithms (Fig-2).

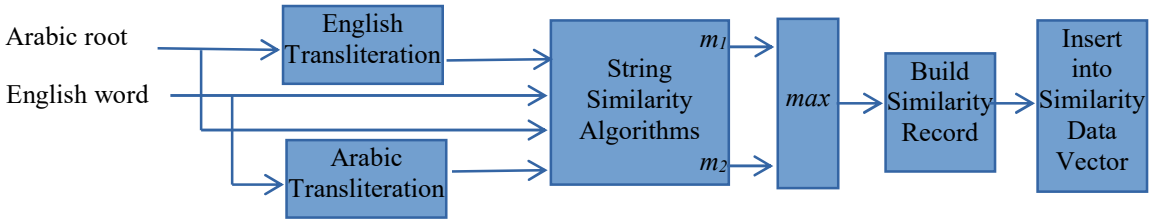


*Figure 2: Computing similarity Data for a Pair of Arabic and English Words*

The larger the similarity value – of these two values - is considered as a measure of the similarity between this pair of words. The main motive behind dual transliteration is to alleviate the effect of factors that may negatively affect transliteration. For example: (1) The script of many words does not exactly match its pronunciation. For example, the word *hour* is actually pronounced as *awar* but transliterated to (هوور). (2) In English, some compositions of letters will have a different pronunciation (e.g. *sh*, *ph, -sion*). (3) The dialects (حركات) of Arabic words are neglected when transliterated to English. For example, (كُحل) is transliterated to (*khl*) and not (*kohl).* (4) The same letter may be pronounced differently in different words depending on the surrounding context. Consider for example the following pairs of words (*can*) and (*cane*), (*kill*) and (*kile*) and the letter *y* in (*eldery)* and (*fly*). (5) Some English letters have varying transliterations. For example, the letter *c* may be pronounced as *s* as in *celery* or as *k* as in *car*.

## 3. TEXT SIMILARITY AGORITHMS

Large number of algorithms may be found in literature for measuring similarity between strings. Each algorithm has its own approach. To avoid bias and shortcoming of a single algorithm, seven different algorithms are executed twice for every pair of words, once for Arabic alphabet and once for English alphabet. These algorithms are: (1) *Dice* [7], (2) *Levenshtein* [8], (3) *Smith Waterman* [9], (4) *Jaro* [10], (5) *Jaro Winkler* [11], (6) *Soundex* - with adjustment for Arabic alphabet [12] and (7) *longest common substring* (*LCS) [13]*. It is important to point out here that some of these algorithms measure the similarity (e.g. *Dice*). This means that the higher the value of the measure, the higher the similarity is. On the other hand, some algorithms measure distance (e.g. *Levenshtein)*. This means that the higher the value of the measure, the lower the similarity is. Therefore, to have a uniform measure of similarity, the distance measures that are given by distance algorithms are transformed to a similarity measure by a reasonable mathematical inversion operation. For example, for *Levenshtein* algorithm, the following formula was used to convert edit distance (*d*) – that is calculated by this algorithm - to a similarity value between 0 and 1.

$$Levenshtein(w_1, w_2) = \{1 - \frac{d}{max(w_1.length, w_2.length)}\}$$

Table-1 shows the similarity values for the Arabic word (كحل) whose translation (*not* English transliteration) to English is *kohl*. The similarity algorithms are run between the word (كحل) against the Arabic transliteration for *every* English word indexed in Oxford dictionary. In addition, the similarity algorithms are run between the transliteration for the word كحل (which will be *khl*) with every word indexed in Oxford dictionary. For every pair of Arabic root and English word, there will be seven similarity values, one for each algorithm.

Journal of Theoretical and Applied Information Technology
15th March 2023. Vol.101. No 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table-1: Similarity Measures Between كحل and kohl*

| Dice | Levenshtein | SmithWaterman | Jaro | JaroWinkler | LCS | Soundex |
|---|---|---|---|---|---|---|
| 0.400 | 0.750 | 0.5 00 | 0.917 | 0.925 | 0.500 | 0.875 |

Each of these similarity measures is the maximum of the two measures: one using Arabic alphabet and another using English alphabet. To illustrate that, let $w_1$ and $w_2$ be an Arabic root and English word (respectively) under consideration. The Levenshtein measure will be:

$$max(Levenshtein(w_1, transliterate(w_2)), Levenshtein(transliterate(w_1),w_2))$$

For the example in Table-1, $w_1$ is the Arabic root (كحل) while $w_2$ is the English word (*kohl),* the value of 0.750 for Levenshtein algorithm will be

$$max(Levenshtein(\text{كحل} , \text{كوهل}), Levenshtein(khl,kohl)) = max(0.5,0.75)= 0.75$$

As mentioned earlier, the similarity between a given Arabic root and all indexed English words in Oxford Dictionary is measured. The problem is that a similarity algorithm will give a similarity value for any pair of words. So, an upper bound for the number of English words to be considered similar to a given Arabic root need to be set. The selected bound was to maintain the top thirty English words with the highest similarity value for *each* of the seven algorithms for a given Arabic root. Table-2 illustrates this for the Arabic root (كحل). Many of the words are common for all algorithms but with different order. Yet, every algorithm may have its own similar words that are not found by other algorithms.

*Table 2: Similarity Measures For The Arabic Word (كحل)*

| | Dice | | Levenshtein | | SmithWaterman | | Jaro | | JaroWinkler | | LCS | | Soundex | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | sikh | 0،4 | **kohl** | 0.750 | col | 0.667 | **kohl** | 0.917 | **kohl** | 0.925 | sikh | 0.500 | kyle | 0.875 |
| 1 | lakh | 0.400 | phlox | 0.571 | sikh | 0.500 | kohlrabi | 0.792 | kohlrabi | 0.8125 | lakh | 0.500 | kola | 0.875 |
| 2 | **kohl** | 0.400 | kyle | 0.5 00 | lakh | 0.500 | c | 0.778 | c | 0.800 | **kohl** | 0.500 | **kohl** | 0.875 |
| 3 | khan | 0،400 | kola | 0،500 | kyle | 0.500 | col | 0.778 | col | 0.800 | khan | 0.500 | kilo | 0.875 |
| 4 | phlox | 0،333 | kilt | 0.500 | kola | 0.500 | kyle | 0.722 | khan | 0.78 | phlox | 0.400 | kill | 0.875 |
| 5 | khaki | | kilo | 0.5 00 | **kohl** | 0.500 | kola | 0.722 | kyle | 0.75 | khaki | 0.400 | keel | 0.875 |
| : | : | : | : | : | : | : | : | : | : | : | : | : | | |
| 29 | gymk hana | 0،222 | excel | 0.429 | skulk | 0.400 | hele | 0.722 | cloy | 0،75 | sh | 0.333 | fly | 0.833 |

For the seven similarity algorithms, the number of similar words for a given Arabic root is bounded by 210 words (seven algorithms with thirty words each). However, the total number of words is far less than this bound because many of the found similar words are common between different algorithms. To ease their handling, these words are categorized into classes according to the number of algorithms that found them. Table-3 shows the words that were found for the Arabic word (كحل). For example, all the seven algorithms found that the English word (*kohl*) to be similar to word (كحل). This table represents the similarity record that is created and inserted into the similarity data vector for the Arabic root (كحل).

*Table-3 The Similarity Record For Arabic Word (كحل) Categorized Into Classes*

| Similar English words | Number of Algorithms |
|---|---|
| kohl | 7 |
| khan | 6 |
| Kyle, kola, kilo, kill, kale | 5 |
| Kilt, kiln, kelt, kelp, keel | 4 |
| Sikh, lakh, phlox, kohlrabi, col, c, | 3 |

| | |
|---|---|
| Khaki, sheikh, cult, cull, colt, cole, cold, cola, col | 2 |
| Richly, phloem, phlegm, highly, gurkha, dahlia, ashlar, schlock, roughly, monthly, kolkhoz, highlight, fleshly, earthly, deathly, chloral, athlete, suchlike, ruthless, pamphlet, highland, gymkhana, exhale, dhal, calx, axle, axil, xerox, phalanx, foxhole, xylem, telex, relax, pixel, keyhole, helix, expel, exile, excel, celt, cell, calm, call, calk, yokel, vocal, skull, skulk | 1 |

## 4. ANAYSIS OF SIMILARITY DATA

### 4.1 Similarity Data Analysis Process

The process of analyzing the similarity data is a manual process that was carried out by the author (Fig-3). Its goal is to identify English words that are very close in pronunciation and meaning to an Arabic one. The user takes the similarity records one by one from the similarity data vector. One similarity record corresponds to one Arabic root. The similarity record structure is very similar to Table-2. He compares the pronunciation of the current Arabic root with the current similar English word. If he finds that the words do not show obvious similarity in pronunciation, this word will be dropped and the next English word will be taken. Otherwise, he will compare the meaning of the two words. If they show very close meaning, he will mark this pair of words as potential English word with Arabic origin.

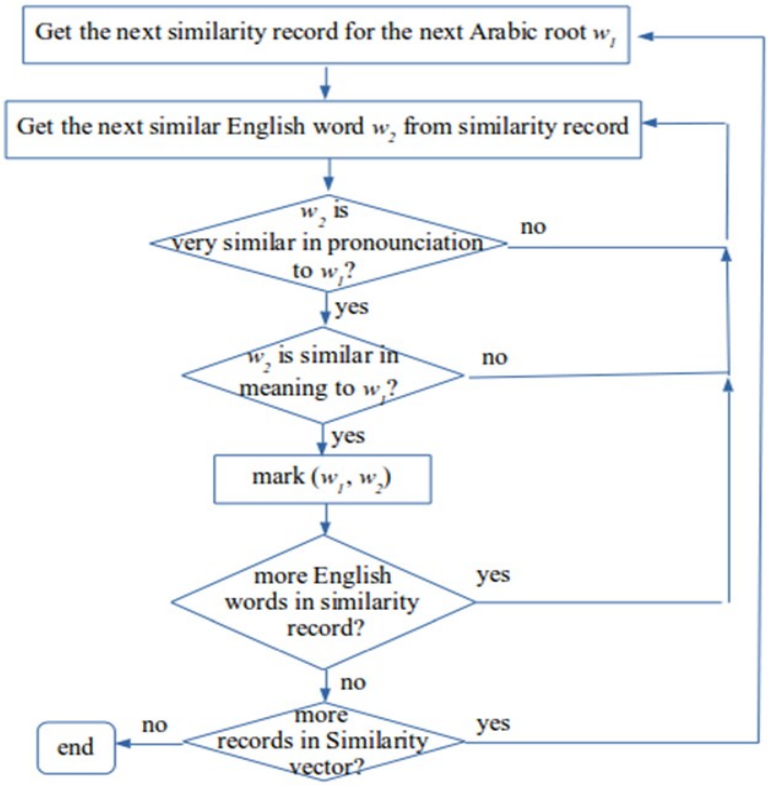


*Figure 3: Similarity Data Analysis Process*

### 4.2 Analysis Process Results

The proposed algorithm was implemented and used to generate similarity data between Arabic roots and English words. This similarity data is analyzed by the author following the process that was presented above. Due to space limitations, only the similarity records of the Arabic roots of the first five chapters of the selected Arabic dictionary [6] were analyzed. These correspond to the roots starting with the Arabic letters (أ, ب, ت, ث, ج). Results are presented in Table-4. Each row shows the detailed information of a pair of Arabic and English words that showed obvious similarity. This table is composed of five columns: (1) Arabic word

along with its English transliteration, (2) the meaning of the Arabic word in English, (3) the similar English word, (4) its meaning and (5) whether if this pair of words is available in *Paradise Dictionary* [2]. The last column shows the contribution of this study. The total number of the found pairs was fifty six pairs. Twenty five of these pairs were NOT found in [2]. It is important to point here that these pairs does not mean that these English words definitely originate from Arabic. All what could be said is that they show very high similarity (in pronunciation and meaning) and hence they *may* have originated from Arabic. Therefore, a further investigation – that is out the scope of this paper - is required. In other words, the value of this work is limited in finding pairs of words that show obvious similarity in pronunciation and in meaning.

*Table 4- Analysis Process Results*

| No | Arabic Root (word) | Arabic Word Meaning | Similar English Word | Meaning in English | in Paradise Dictionary? |
|---|---|---|---|---|---|
| 1 | أرن (ar n) | A place used as a shelter | barn | A large farm building used for storing grain, hay, or straw or for housing livestock. | no |
| 2 | ألب (alaba) | to collect | album | A blank book for the insertion of photographs, stamps, or pictures. | no |
| 3 | أطر (atara) | to surround something | attire | To be dressed with clothes of a specified kind. | no |
| 4 | أطم (atam) | to prison or surround | atom | The smallest particle of a chemical element that can exist | no |
| 5 | أكّ (ak'ka) | to have hardship from heat or something else | ache | Suffer from a continuous dull pain. | yes |
| 6 | أير (ayara) | Wind | Air | Light wind | yes |
| 7 | أيّ (ay'ya) | to look | eye | Sight | yes |
| 8 | بطش (batasha) | Kill in cruel manner | butch | Slang masculine | yes |
| 9 | بثّ (bath'tha) | to distribute, spread, show | booth | Small temporary structure used esp. as a market stall | no |
| 10 | بجّ (baj'ja) | Full body, extremely watering camels to full | Baggy | Hanging loosely, baggy trousers | no |
| 11 | بر (burr) | wheat | bur | A prickly clinging seed-case or flower head or any plant having these. | no |
| 12 | بدل (badala) | to replace something for another | Peddle | A sell (goods) as a peddler | no |
| 13 | برج (baraja) | Talk arrogantly | brag | Talk boastfully | yes (برق) |
| 14 | برك (baraka) | to sit (for camel) | park | …, park oneself colloq. Sit down | yes |

| | | | | | |
|---|---|---|---|---|---|
| 15 | بغي (baghy) | aggression | bogey | Evil or mischievous spirit | yes |
| 16 | بقّ (baqqa) | Small things, type of insects | bug | Any of various insects with mouthparts modified for piercing and sucking. | yes |
| 17 | بكّ (bakka) | to crowd | pack | Crowd or cram | no |
| 18 | بلع (bala'a) | to swallow | bugle | Irregular swelling. | no |
| 19 | بهم (bahama) | The thing that is not known how to deal with | bohemian | Socially unconventional person: not behaving in the same way as most other people in society. | yes |
| 20 | بور (boor) | Bad (loser) person | boor | Ill-mannered person | yes |
| 21 | تبن (tobn) | A big cup that may serve twenty persons | tun | Large beer or wine cask | no |
| 22 | توأم (taw'am) | twin | twin | Each of a closely related or associated pair | yes |
| 23 | ترص (tarasa) | Anything that is tightened, to support | Truss<br>tress<br>terse | - Framework supporting a roof, bridge<br>- Fong lock of human hair<br>- Brief, concise | no<br>no<br>yes |
| 24 | تقن (taqana) | to work skillfully | technology | Knowledge or use of the mechanical arts and applied sciences | yes |
| 25 | تلع (tala'a) | To extend, to lengthen | tall | of more than average height | Yes (طول) |
| 26 | تلّ (tal'la) | Erect vertically | tall | of more than average height | no |
| 27 | باص (basa) | to pass in a hurry | pass | To pass | yes |
| 28 | بيع (baya'a) | to sell | buy | To obtain for money | yes |
| 29 | بين (bayana) | to be Exposed, open land | open | Not closed, allowing access, .. exposed … expanded … | no |
| 30 | ثخن (thakhana) | to make/be dense (liquids) | thicken | To make thick (dense) | yes |
| 31 | جأب (ja'aba) | to gain, to work | job | Position in, or piece of paid employment | no |
| 32 | جبل (jabala) | to gather something with elevation | gable | Triangular upper part of a wall at the end of a ridged roof | yes |
| 33 | جبر (jabara) | reunion of broken parts[1] | algebra | Branch of mathematics that uses letters etc. to represent numbers | yes |

[1]The word algebra comes from the Arabic: الجبر, romanized: al-jabr, lit. 'reunion of broken parts,

| | | | | | |
|---|---|---|---|---|---|
| | | | | and quantities. | |
| 34 | جثّ (jath'tha) | The gathering of something | gather | Bring or come together | no |
| 35 | جحظ (jahaza) | The eye to pop out | gaze | Look fixedly | yes (خزر) |
| 36 | جرّ (jar'ra) | to poll something | garrote | Execute or kill by strangulation, esp. with a wire collar | no |
| 37 | جرثم (jarthama) | to cut a piece of land and stuck with it | germ | Rudiment of an animal or plant in seed (wheat germ) | Yes (الأرومة) |
| 38 | جرج (jaraja) | The used passage (way) | gorge | Narrow opening between hills. | no |
| 39 | جرح (jaraha) | to wound (open the skin) | gore | Bloodshed and clotted pierce with a horn, tusk, etc. | no |
| 40 | جرد (jarada) | to show something without a cover or shield | gird | - Encircle, attach, or secure, with a belt or band<br>- Enclose or encircle. | no |
| 41 | جرش (jarasha) | To pound into crumples | crush | Cease and crumple | yes |
| 42 | جرع (jara'a) | To drink a little | grog | Spirits (originally rum) mixed with water. | no |
| 43 | جرل (jarala) | -Rocks جرول (jarwal) | gravel | mixture of coarse sand and small stones | no |
| 44 | جرن (jarana) | (jorn)<br>- place were grains are collected to be cleared from straw | grain | Fruit or seed of a cereal, wheat or any allied grass used as food | yes |
| 45 | جزل (jazala) | - to be alot | guzzle | To eat and drink greedily | no |
| 46 | جشر (jashara) | - to spread and become clearly seen | gush<br>gusher | A rapid and plentiful stream or burst of something.<br>- an effusive person | no |
| 47 | جلد (jalada) | - skin | gild | - To cover thinly with gold | no |
| 48 | جلط (jalata) | - to get rid from something | Jilt | - Abruptly reject or abandon | yes |
| 49 | جلو (jalawa) | - to get exposed and apparent | glow | - Emit light, shine, to show strong emotion | yes |
| 50 | جمن (jamana) | - pearls (جمان) | gem | - Precious stone | yes |
| 51 | (جند) janada | - army | gendarme | - Army | yes |
| 52 | جنس | - a type or class of | genus | - Kind or class | yes |

bonesetting' from the title of the early 9th century book *al-jabr wa l-muqābala* "The Science of Restoring and Balancing" by the Persian mathematician and astronomer al-Khwarizmi [14].

| | | | | | |
|---|---|---|---|---|---|
| | (jins) | something | | | |
| 53 | جنّ (jan'na) | - to hide<br>- Jinnee | Jinnee | - Spirit in human or animal form having power over people | yes |
| 54 | جني (janaya) | - to collect the fruit<br>- to gain | gain | - To acquire as profit or earn | yes |
| 55 | جود (jawada) | - to be forgiving, to give generously | good | - Good | yes |
| 56 | جيب (jayaba) | - to make a whole | jap | - Poke roughly or quickly with something sharp or pointing object | no |

## 5. CONCLUSION

In this paper, string similarity algorithms were used to measure the similarity between Arabic roots and English words. This helped in identifying English words that *may* have originated from Arabic. The main contribution of this paper is utilizing computation to serve this purely linguistic issue. Another contribution is the identification of a number of English words that showed obvious similarity with Arabic words and yet were not listed in *Paradise* dictionary. However, this study covered less than the sixth of the Arabic roots that are listed in the selected Arabic dictionary. As future work, the rest of the Arabic roots may be studied. Therefore, more English words that potentially originate from Arabic may be identified. In addition to that, the analysis of the similarity data process was completely manual. A friendly graphical user interface may be developed - that is connected with dictionaries - to ease the analysis of the similarity data. This interface may automate the display of this data along with the meaning of words without manual return to dictionaries.

## REFERENCES

[1] DICTIONARY, Oxford English. Oxford English Dictionary. Simpson, Ja & Weiner, Esc, 1989.

[2] S. Abu Ghoush, "10000 English Word from Arabic", Kuwait. Printing Agency, 1977.

[3] H. Darwish, "Arabic loan words in English language", *Journal of Humanities and Social Science*, Vol. 20, No. 7, 2015, pp. 105-109.

[4] M. Yacoub, "Do Ten Thousand Arabic Loanwords Truly Exist In English?" *International Journal of Interdisciplinary Research and Innovations*, Vol. 3, No. 1, 2015, pp. 102-106.

[5] G. Cannon, A. Kaye, "The Arabic contributions to the English language: an historical dictionary", *Otto Harrassowitz Verlag*, 1994.

[6] IbnFaris, "*Mojam Maqayees Al-lugha"* [The measures of The Language dictionary], Dar Al-hadeeth, Cairo, 2008.

[7] L. Dice, "Measures of the amount of ecologic association between species", *Ecology,* Vol. 26, No. 3, 1945, pp. 297-302.

[8] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet physics doklady*, Vol. 10, No. 8, 1966, pp. 707-710.

[9] S. Temple, M. Waterman, "Identification of common molecular subsequences", *Journal of molecular biology,* Vol. 147, No. 1, 1981, pp. 195-197.

[10] M. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida", *Journal of the American Statistical Association*, Vol. 84, No. 406, 1989, pp. 414-420.

[11] W. Winkler, "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage", 1990.

[12] D. Knuth, "The Art of Computer Programming: Sorting and Searching", Addison-Wesley, 1998, pp. 75-80.

[13] D. Gusfield, "Algorithms on stings, trees, and sequences: Computer science and computational biology", Acm Sigact News, Vol. 28, No. 4, 1997, pp. 41-60.

[14] https://en.wikipedia.org/wiki/Algebra, 19/2/2023.