INTERNATIONALJOURNAL OF ADAVNCED STUDIES
IN COMPUTER SCIENCE AND ENGINEERING 01/31/2017
IJASCSE VOLUME 6 ISSUE 01, 2017

# Auto-Extracting Method of Cognates Words in Arabic and English Languages

Hamad A. Alreshidi
Instructional technology dept. education college
University of Hail , Hail , KSA

Kawther A. Al-dhlan
CSSE dept., College of Computer science& engineering ,
University of Hail , Hail , KSA

***Abstract***

**For a long time computers have been used for language teaching and learning. Unfortunately, little has been done in the field of computerized Arabic language teaching and learning. Furthermore, there is a lack of research into the impact of applying the knowledge of cognates (words similar in meaning and pronunciation) in an Arabic web-based learning system for English-speaking learners. This study investigates the effects of word cognates (words that have same pronunciation and meaning) used in Computer-Assisted Language Learning systems. In addition, it studies the enhancement of a language learning using Computer-Assisted Language Learning systems. The proposed system is able to identify cognates automatically in any of the two languages (Arabic and English languages) and presents them in a list. For this purpose, the research developed a web-based system for use as an Arabic language learning tool for English-speaking learners. The system automatically identifies cognates in the two languages using one of two functions: levenshtein() and similar_text(). The system displays the cognates in a list (table form) and highlights the cognates in the lessons. Using cognates will improve learners' capability of developing the target language (in this research English, if the learner's tongue language is Arabic, or Arabic if the learner's tongue language is English).**

***Keywords: cognates; Arabic; English; languages; CALL.***

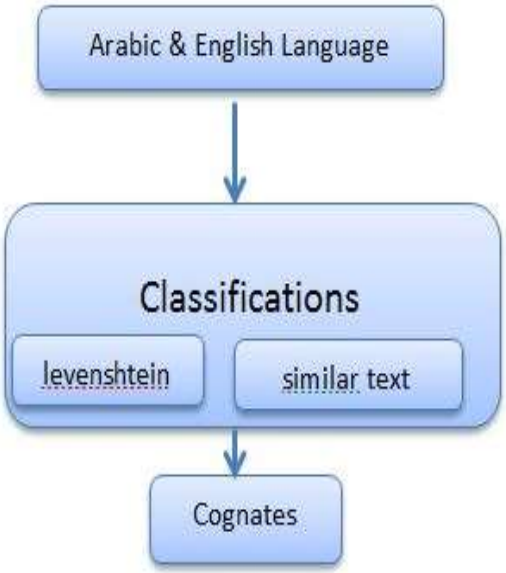


Fig. 1 Schematic Diagram of the Research Implemented

## INTRODUCTION

This section gives a brief background on the research worked carried out by this study. Fig. 1 shows the schematic diagram for the implemented research. The section focus on the important of Arabic language with different aspects.

Arabic language is one of the most popular languages (see figure 2) that people are spoken and willing to learn. The Arabic language is the fifth most commonly spoken native language in the world. In addition, Arabic is the official language of 20 countries, as well as an official language of the Organization of Islamic Conference, the Arab League, the United Nations and the African Union (Sirajudeen & Adebisi, 2012).


Figure 2 English words with some Arabic origins (Oxford Dictionaries, 2014)

Arabic has had a great influence on other languages, especially in vocabulary. The influence of Arabic has been most profound in those countries dominated by Islam or Islamic power.

Arabic language is a major source of vocabulary for languages as diverse as Amharic, Bengali, Hindi, Indonesian, Kazakh, Kurdish, Kyrgyz, Malay, Pashto, Persian, Punjabi, Sindhi, Somali, Swahili, Tigrinya, Turkish, Turkmen, Urdu, Uyghur, Uzbek as well as other languages in countries where these languages are spoken. For example, the Arabic word for book (كتاب kitāb) is used in most of the languages listed, (exceptions are Spanish, Catalan and Portuguese, which use the Latin-derived words "libro," "llibre" and "livro"). Spanish language (as a example) has the largest Arabic influenced vocabulary outside the Islamic world due to Muslim rule in the Iberian Peninsula from 711 until 1492 known as Al-Andalus.

Unfortunately, the application of computer technology in the teaching of the Arabic language is still at the development stage (Ghalib, Hassan, Sahrir, Daud, Mohin & Ayu, 2014). Suliman, Mukhtar and Sahrir (2013) state that the challenge is the lack of Arabic CALL resources in the Arabic language.

To enhance learning of the Arabic language, the knowledge of cognates could be used (cognates are words similar in meaning and pronunciation). Using loanwords to help with learning of a language (Arabic) is a very effective vocabulary expansion strategy (Nation, 2003). Furthermore, bringing cognates to the attention of learners could help

Unfortunately, learning the Arabic language as a second language is considered to be difficult among learners (see Table 1). Hence, it is recommended that instructors use technology and to enhance their teaching of the Arabic language, as use of technology and multimedia motivates learners and enhances their performance in the classroom (Faryadi, 2007). However, using technology for Arabic language learning is not that simple. One of the challenges facing Arabic Computer-Assisted Language Learning (CALL) is that Arabic is a non-Roman orthography language (Bush, 2004).

| ARABIC | ENGLISH |
|---|---|
| Semitic Language Family | Germanic language family |
| Written in cursive script | Most often printed |
| Read from right to left | Read from left to write |
| 28 letters in the alphabet | 26 letters in the alphabet |
| Letters change based on their position in a sentence | Letters are fixed |
| There are only 3 vowels /a/, /u/, /i/ | There are 5 vowels (not including y which sometimes behaves as a vowel) |
| No capitalization | Capitalization is used at the beginning of a sentence and for proper nouns |
| Limited punctuation - commas are used loosely | Punctuation is essential and indicates the completion of thoughts, emphasis and inflection |
| Phonetic spelling | Spelling is not entirely phonetic |
| Verb first sentences | Verb last sentences |
| Three consonant root language | |

Table 1: What are the key differences between Arabic and English (Bray, 2011)

in extending their vocabularies and improving their comprehension (Goldenberg, 2008).

Thus, the importance of this research work is in finding a creative language learning environment for learning Arabic language. This study aims to help learner overcome the challenges and difficulties of learning the Arabic language by employing a combination of the CALL methodology and the use of knowledge of a cognate strategy for learning the language.

It is hoped the research will familiarize English-speaking students with cognates in Arabic and English, helping them to understand the link between these two languages and thus aiding them in learning the Arabic language. Moreover, the research aims

to enhance current Arabic e-learning systems by adding a strategy of cognate awareness in self-studying.

### 1.1 Computer-Assisted Language Learning

The computer is used as a tool throughout education, including in language learning. In the language area, as identified by Abu Naba'ih, Hussain, Al-omari, and Shdeifat (2009), Computer-Assisted Language Learning (CALL) is the approach of using a computer as a tool for learning and teaching a language.

CALL makes learning easier and more flexible. Many websites and software packages provide plenty of foreign language learning materials to be easily accessed by the learner and the educator. The plentiful resources and educational equipment that are provided by CALL gives this learning methodology the potential to replace the traditional learning methodology, since it works in more flexible environments and has better options in relation to the learner/teacher needs, interests and abilities (Uzun, 2012).

The power of multimedia supports CALL systems because it enables integration with graphics, video, and sound (Laghos & Zaphiris, 2005).

### 1.2 Learning Cognates

Cognates are words that have a similar meaning, spelling and pronunciation in two languages (Kondrak, 2009). However, similarities among languages are not necessary due to borrowing or the origin of one language. Campbell and Poser (2008) argue that some scholars identifying the similarities among languages are only due to inheritance from a common ancestor.

They mention other reasons for the existence of similarities among languages, such as by accident (chance, coincidence), onomatopoeia, universals and typologically commonplace traits. Hara (2011) states that learning cognates can help facilitate learning of a foreign or second language (L2) because cognates preserve the linguistic information of the first language (L1). There is widespread agreement that using cognates helps in facilitating the learning process, especially if the cognates in L1 are loanwords from L2 or the target language (TL), and if these cognates are words of relatively high frequency (Ringbom, 2007).

Otwinowska and Kasztelanic (2012) conducted a study to examine how using Polish cognate vocabulary exercises can affect the learning of learners of English and what their attitudes were toward this strategy. A quasi experimental design and qualitative methods were used to gather learners' opinions and attitudes. A total of 14 students participated and they were randomly divided into two groups: an experimental group and a control group.

The findings showed that raising awareness of cognates helped the teenagers to recognize cognates more effectively while reading an English text. The majority found that the exercises helped them notice similarities between Polish and English. The study highlights how cognate vocabulary can be important for the speed of language acquisition.

A study conducted by Ibrahim (2006) on Arabic speakers and Hebrew speakers to examine whether cognates in one's first language are an advantage to second language students' learning concluded that similarities between languages reflected by cognates' relationships can influence linguistic bilingual performance. Again, Lapo claimed that knowledge of the similarities and differences between Spanish and English is crucial in regards to establishing connections between the two languages and in facilitating positive transfer from Spanish to English.

Similarly, in another study on the acquisition of English vocabulary by Chinese learners it was found that foreign words are remembered by being linked to a keyword, a sound-alike native word (the acoustic link), or an interactive image that involves both the foreign word and the native word (the imagery link) (Zhang, 2005). Specifically, lessons that incorporate cognates (e.g. individual (English) and individuo (Spanish)) have been found to be effective in expanding learners' English vocabulary development and aids in comprehension (Proctor, Dalton, & Grisham, 2007).

The findings of yet another study suggest that while literacy in Spanish can provide students with access both to orthographic as well as phonological sources of information about cognate relationships, it is possible for students to draw connections between cognate pairs on the basis of sound alone, so that students who are not literate but orally proficient in Spanish are likely to benefit from instruction in cognate awareness as well as those who are literate in Spanish (Dressler, 2000).

### 1.3 Transliteration from Arabic to English

Different approaches and methods for the transliteration of Arabic exist. They vary in the way that they address the inherent problems of rendering written and spoken Arabic in the Latin script. Examples of such problems are the symbols for Arabic phonemes that do not exist in English or other European languages; the means of representing the Arabic definite article, which is always spelled the same way in written Arabic but has numerous pronunciations in the spoken language depending on context; and the representation of short vowels (usually i u or e o, accounting for variations such as Muslim/Moslem or Mohammed/Muhammad/Mohamed).

These are sample of general transliteration guidelines set by International Journal of Middle East Studies (IJMES, 2016):

1. If an English term exists for a word, use it.
2. All technical terms from languages written in non-Latin alphabets must be italicized and fully transliterated with diacritical marks (macrons and dots), e.g., ʿashāʾ. A technical term is defined as a word for which there is no English equivalent and that is not found in Merriam–Webster's Collegiate Dictionary, or a multi-word phrase, excluding names and titles as detailed in #4 below. Diacritical marks, as well as the letters ʿayn and hamza, should be inserted using a Unicode font.
3. Words found in Merriam–Webster's should be spelled as they appear there and not treated as technical terms. They should have no diacritics, nor should they be italicized—for example, mufti, jihad, shaykh. Some words can have an exception that preserve ʿayn and hamza, for example, Qurʾan, shariʿa, ʿulamaʾ, and Kaʿba.
4. Diacritics should not be added to personal names, place names, names of political parties and organizations, or titles of books and articles. However, ʿayn and hamza should be preserved in all these cases (except for initial hamza, which is dropped). Do not italicize transliterated proper names, including titles of organizations, and do follow English capitalization rules: for example, al-Madina al-Munawwara.
5. Arabic names of prominent political or cultural figures are spelled without diacritics e.g., Jamal ʿAbd al-Nasir. Names of living individuals may be spelled according to their preferred English spelling.
6. Place names with accepted English spellings should be spelled in accordance with English norms, for example, Baalbek, Damascus.
7. Follow English capitalization rules for transliterated titles. Capitalize all major terms, but not articles, prefixes, coordinating conjunctions, or prepositions (even when joined to pronouns). Use italics to indicate a book, newspaper, or periodical. Do not include diacritical marks but do preserve ʿayn and hamza. Ex. Faysal al-Tafriqa bayn al-Islam wa-l-Zandaqa; al-Nur al-Safirʿan Akhbar al-Qarn al-ʿAshir.
8. Avoid Anglicized plurals on fully transliterated words if possible: for example, fuqahā' , not faqīhs. Exceptions may be made if there is a good reason for it, such as when comparing numbered quantities or currencies. Anglicized plurals may be used on words that are found in Merriam-Webster's and thus not fully transliterated, e.g., muftis.
9. When in doubt, follow the spelling of the term in the script of the original language, not its oral pronunciation. There are only a few exceptions (e.g., iḍāfa constructions).

## I. PROBLEM STATEMENT

This section focus on the problem statement carried out by this research. Little has been done in the field of computerized Arabic language teaching and learning. Very few researchers have studied the use of cognates by teachers (traditional learning) as a tool for teaching the Arabic language. Some previous studies have focused on language learning assisted by the computer and others have focused on language learning strategies.

Some have even created algorithms or improved old ones to automatically identify cognates. However, none of them have considered a combination of CALL, cognate transfer and highlighting cognates.

Despite the existence of these previous studies, no study has developed an e-learning system for language learning using the knowledge of cognates. There is a lack of research into the impact of applying the knowledge of cognates in an Arabic web-based learning system for English speakers.

## II. OBJECTIVES

- The objectives of this study are as follows:
- To design a system (software) to identify the similar cognates between English and Arabic languages.
- To build a database of the similar cognates between English and Arabic languages.

## III. TARGET USERS

Many educational sectors and individuals may benefit from the proposed system and find their interest and desire to purchase and use such as

- Individual languages learners
- Languages institutes
- Online language learning industries
- Languages tutors
- Schools and Universities

## IV. RESEARCH ACTIVITIES

This section will show the most important activates to be followed by this study, we will focus on 8 activities as shown below:

- Study on related works.
- Collect data from dictionaries in both languages (Arabic and English) and include them in database.
- Develop an algorithm to identify similar cognates
- Design a system to apply the algorithm
- Insert the database to the designed system
- Test the system and correct the errors
- Publish a paper about the proposed work
- Finalize the project and submit the final report

## V. RELATED WORKS

This section will discuss some previous studies related to this research. There are very limited researches done about using computer learning system using word cognates of Arabic and English languages. On the other side, there are many studies using word cognates of other languages.

In (Malmasi and Dras, 2015), they described an approach to automatic cognate identification in mono-lingual texts using machine translation. they proposed an approach that takes an input text in a source language and uses statistical machine translation to create a word-aligned parallel text in the target language. A robust measure of string distance, the Jaro-Winkler distance in this case, was then applied to the pairs of aligned words to detect potential cognates.

Based on the assumption that the linguistic of words changes follow certain rules, (Ciobanu and Dinu, 2014) proposed a method for automatically detecting pairs of cognates employing an orthographic alignment method which proved relevant for sequence alignment in computational biology. They uses aligned subsequences as features for machine learning algorithms in order to infer rules for linguistic changes undergone by words when entering new languages and to discriminate between cognates and non-cognates.

Another study (Biller, 2010) explored the role of phonology in the processing of Arabic words in native English speakers learning Arabic. She utilized phonologically embedded English words within Arabic words at three different points in the word, beginning, middle, and end, to determine the role of phonology separate from that of semantics. The results generally support the importance of phonology in the mental processing of multiple languages, which can be combined with other findings in language research to supplement language learning programs.

A loosely-supervised method (Goldberg and Elhadad, 2007) for context-free identification was presented of transliterated foreign names and borrowed words in Hebrew text. The method is purely statistical and does not require the use of any lexicons or linguistic analysis tool for the source languages (Hebrew, in this study). It also does not require any manually annotated data for training. They report precision/recall results of 80/82 for a corpus of 4044 unique words, containing 368 foreign words.

(Sherif and Kondrak 2007) used seed examples and a sentence aligned English/Arabic text to jointly learn a bilingual string distance function and extract transliterated pairs. While this work aims at complete alignment, their task was only the identification of transliterated candidates. Identification of transliteration candidates can help full alignment by relaxing the need for aligned text.

In this study (Saadane, et al. 2009), they focus on the use of Arabic transliteration to improve the results of a linguistics-based word alignment approach from parallel text corpora. This approach uses, on the one hand, a bilingual lexicon, named entities, cognates and grammatical tags to align single words, and on the other hand, syntactic dependency relations to align compound words. They have evaluated the word aligner integrating Arabic transliteration using two methods: A manual

evaluation of the alignment quality and an evaluation of the impact of this alignment on the translation quality by using the Moses statistical machine translation system. The obtained results show that Arabic transliteration improves the quality of both alignment and translation.

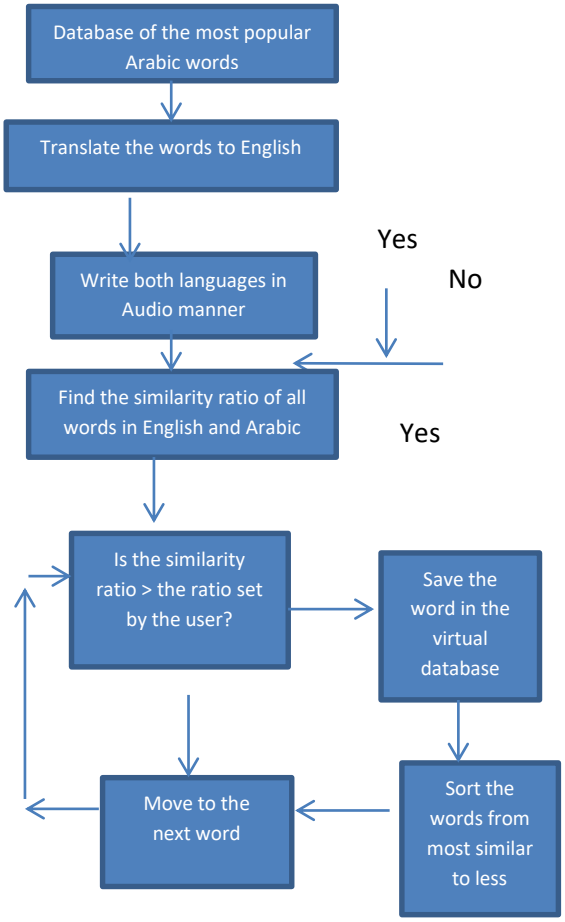


Figure 3: Research methodology

i. Designing a CALL system using two of the existing functions in the computer programming languages: similar_text() and levenshtein.
ii. Developing a prototype of Arabic Web-based CALL system using the knowledge of cognates
iii. Evaluation of the prototype using a quantitative method and descriptive analysis

- **Functions used to determine similarity**

Determining similarity between two strings is called the 'Levenshtein distance'. The Example Based Machine Translation (EBMT) (Aaron and Sforza, 2012) is a form of automated translation that uses a large corpus of previously-translated example sentences to create a translation for a new sentence. Typically the system does not have in its corpus the entire sentence to be translated. Instead, the system matches words and small phrases and stitches them together with the help of a target language model. Although EBMT uses statistical methods, it differs from Statistical Machine Translation (SMT). EBMT consults its corpus of translations at runtime, whereas SMT prerocesses the corpus to calculate the probability of a word or phrase occurring as translations, but then uses only these probabilities at runtime.

## VI. METHODOLOGY

This section will discuss the methodology of this research by steps and phases as follows:Studying the concept of learning cognates and employing the approach of Computer-Assisted Language Learning (CALL) in facilitating the language learning particular for Arabic throughout reviewing the previous relevant studies. Levenshtein distance is named after mathematician Vladimir Levenshtein, who developed the algorithm in 1965 (Gosselin, Kokoska, & Easterbrook, 2011). Programmatically, there are many functions which are used to determine the similarity between two strings. Two of those functions are similar_text() function and levenshtein() function.

- **Function One (similar_text())**

The similar_text() function returns the number of letters that two strings share (Gosselin et al., 2011). It accepts two string arguments representing the values that need to be compared. For example, applying the function on the 'car' and 'can' string arguments will return the result '2' because there are two characters that the two strings have in common. It allows an alternative third argument which is used to store the commonality between the two strings as a percentage (Tatroe, MacIntyre, & Lerdorf, 2013).

To store these values it should be declared in its proper syntax. Programming language syntax refers to the form and structure of the programs and how they look (Webber, 2010). The syntax of this function is as follows:

similar_text(string1,string2,percent)

- **Applying the Function in the System**

In the system there are three (3) parameters (arguments), as follows:

1. The first string
– which is the Arabic transliteration stored in the dictionary's database.
2. The second string
– which is the English meaning of the Arabic first string stored in the dictionary database.
3. Percent
– which is a float argument passed in the function to calculate the similarity in percent between the first string (Arabic transliteration) and the second string (Englih).

All data stored in the dictionary database will be passed in the function. If the similarity in percent between the 'first string' (Arabic transliteration) and the 'second string' (English) is equal or greater (= >) than the given 'percent' value, the function will return these values:

1. The origin Arabic word that is stored in the dictionary database and which is represented as Latin letters (transliteration) in the 'first string'.
2. The meaning of it in English, the 'second string'.

- **Algorithm of the Function**

Levitin (2012: 3) defines an algorithm as "a sequence of unambiguous instructions for solving a problem, i.e., for obtaining the required output for any legitimate input in a finite amount of time". Levitin states that an algorithm can be described in pseudocode. He defines a pseudocode as "a mixture of a natural language and programming language like constructs. Pseudocode is usually more precise than natural language, and its usage often yields more succinct algorithm descriptions" (Levitin, 2012: 12-13).

He also indicates another method to specify an algorithm using a flowchart. He defines a flowchart as "a collection of connected geometric shapes containing descriptions of the algorithm's steps" (Levitin, 2012: 13).

In this section both the pseudocode (see Figure 3) and flowchart (see Figure 4) of the similar_text() function are described.

Numerous variables are used in the figures and they can be described as follows:

**i. ArabicWord:**
The origin Arabic word that is stored in the dictionary's database.

**ii. transliteration:**
The Arabic transliteration stored in the dictionary's database.

**iii. EnglishWord:**
The English meaning of the Arabic word that is stored in the dictionary's database.

**iv. varPercent:**
The number entered by the user to choose the similarity percentage between the Arabic and English word to be displayed.

```
Step 1: Start
Step 2: Declare variables ArabicWord,
transliteration, EnglishWord, varPercent.
Step 3: Read variables transliteration,
EnglishWord, and percent.
Step 4: calculates the similarity between
'transliteration' and 'EnglishWord' in percent
Step 5: If percent >= varPercent %
        Display ArabicWord and EnglishWord.
Else
        Stop
Step 6: Stop
```

Figure 4: Pseudocode of similar_text function

Figure 5: Flowchart of similar_textfunction

- **Code of the Function**

The PHP scripting language used in coding the interfaces. The code runs "SQL SELECT" query to find all matched words based on the similarity percentage entered by the user. The list of words will be displayed in the interface. The code is as follows (Shehab et al., 2014):

```
<?PHP
```

```
$varPercent = $_POST['txtPercent']; // txtPercent is the input field
that would the user        // enter the similarity percentage he/she
prefers.
$SQL = "SELECT * FROM englisharab ";// englisharab is the
table where the        // dictionary is stored $result=
mysqli_query($db_handle,$SQL);
while( $db_field = mysqli_fetch_assoc($result) ) {
$transliteration = $db_field[' transliteration '];
$EnglishWord = $db_field[' EnglishWord '];
$ArabicWord = $db_field[' ArabicWord '];
similar_text($transliteration, $EnglishWord, $percent);
if ($percent >=$varPercent){
echo $ArabicWord;
echo $EnglishWord;}}
<?php
```

- **Function Two (levenshtein())**

The levenshtein() function returns the number of characters that need to change for two strings to be the same (Gosselin et al., 2011). It accepts two string arguments representing the values that need to be compared. For example, applying the function on the 'car' and 'can' string arguments will return the result '1' because there is only one character that needs to change for the two arguments to be the same.

Alternatively, three values could be passed to the function to individually weight insertions, deletions, and replacements; for instance, to compare a word against a contraction (Tatroe et al., 2013). The syntax of this function is as follows:

**levenshtein(string1,string2,insert,replace,delete)**

- **Applying the Function in the System**

In the system there are two (2) parameters or arguments:

1. The first string

– which is the Arabic transliteration stored in the dictionary database.

2. The second string

– which is the English meaning of the Arabic first string stored in the dictionary database.

In the Levenshtein function, the percent value is not passed in the argument as in similartext() function. Therefore, to find the similarity between the two strings the following steps can be followed:

1. Return the Levenshtein distance between the two argument strings, which is the distance in the number of characters that need to be replaced, inserted or deleted to transform the first string, 'Arabic transliteration', to the second string, 'English word'.

2. Calculate the length of the first string and second string to find the longer one.

3. Perform the operation ((the longer string-Levenshtein distance)/the longer word)* 100, where the final result is the percentage value (Hussain, 2012).

All data stored in the dictionary database will be passed in the function. If the similarity in percent between the 'first string' (Arabic transliteration) and the 'second string' (English) is equal or greater (= >) than the given 'percent' value, the function will return these values:

1. The origin Arabic word that is stored in the dictionary database and which is represented as Latin letters (transliteration) in the 'first string'

2. The meaning of it in English, 'the second string'.

- **Algorithm of the Function**

This section explores the pseudocode (see Figure 5) and flowchart (see Figure 6) to describe the algorithm of the function (refer to section 5.1.1.2 for the definition of these three terms). Numerous variables are used in the figures and they can be described as follows:

**i. ArabicWord:**

the origin Arabic word that is stored in the dictionary's database.

**ii. transliteration:**

the Arabic transliteration stored in the dictionary's database.

**iii. EnglishWord:**

the English meaning of the Arabic word that is stored in the dictionary's database.

**iv. varPercent:**

the number entered by the user to choose the similarity percentage between the Arabic and English word to be displayed.

**v. trLeng:**

the length of the transliteration.

**vi. enLeng:**

the length of the English Word.

**vii. Levenshtein:**

Stores the result of levenshtein() function that is applied to the transliteration and the English word.

See Figure 7 for examples of finding the similarity percentage using levenshtein() function.

```
Step 1: Start
Step 2: Declare variables ArabicWord, transliteration, EnglishWord,
trLeng, enLeng, Levenshtein, varPercent and percent.
Step 3: Read variables transliteration, EnglishWord, and varPercent.
Step 4: Calculate trLeng = the length of 'transliteration' and enLeng = the
length of 'EnglishWord'.
Step 5: Calculate Levenshtein = the characters needed to be changed fr
'transliteration' and 'EnglishWord' to be the same.
Step 6: If trLeng > enLeng
        Calculate percent = ((trLeng – Levenshtein) / trLeng) * 100
Else
        Calculate percent = ((enLeng – Levenshtein) / enLeng) * 100
Step 7: if percent > varPercent
```

```
        Display ArabicWord and EnglishWord.
Else
        Stop
Step 8: Stop.
```

Figure 6: Pseudocode of levenshtein function

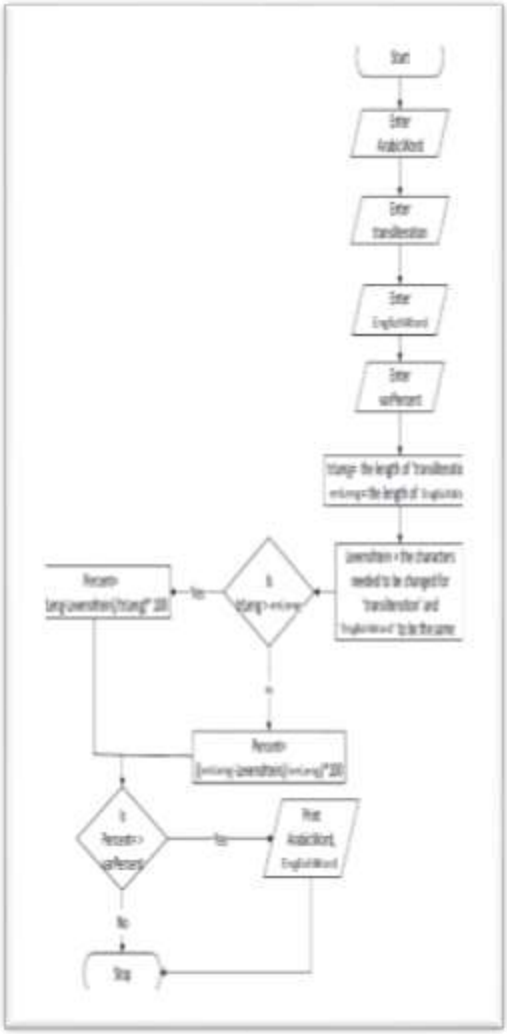

Figure 7: Flowchart of levenshtein function

1) (بنان, banana)
b a n a n (transliteration of بنان)
b a n a n a (the English word)
Levenshtein = 1 (an 'a' is added at the end)
Similarity percentage = ((the length of longest string – levenshtein) / the length of longest string) * 100
= ((6 – 1) / 6) * 100 ≈ 83%

2) (كوب, cup)
c u p (transliteration of كوب)
c u p (the English word)
levenshtein = 0 (no changes is needed)
Similarity percentage = ((the length of longest string – levenshtein) / the length of longest string) * 100
= ((3 – 0) / 3) * 100 = 100%

3) (كتاب, book)
k i t a b (transliteration of كتاب)
b o o k (the English word)
Levenshtein = 5 (all the letters of 'kitab' need to replaced or deleted)
Similarity percentage = ((the length of longest string – levenshtein) / the length of longest string) * 100
= ((5 – 5) / 5) * 100 = 0%

Figure 8: Examples of finding the similarity percentage using levenshtein function

- **Code of the Function**

The code in PHP used to run the function is as follows (Shehab et al., 2014):

```
<?PHP$varPercent=$_POST['txtPercent
'];
$SQL = "SELECT * FROM englisharab
";$result = mysqli_query($db_handle,$SQL);
while ( $db_field =
mysqli_fetch_assoc($result) ) {
$transliteration = $db_field[' transliteration
'];$EnglishWord = $db_field[' EnglishWord '];
$ArabicWord = $db_field[' ArabicWord
];$trLeng=strlen($transliteration);
$enLeng=strlen($EnglishWord);
$lev= levenshtein($transliteration,$
EnglishWord);if
($trLeng>$enLeng){$percent= (($trLeng-
$lev)/$trLeng)*100;}
else{$percent= (($enLeng-
$lev)/$enLeng)*100;}
if ($percent >=$varPercent)
echo $ArabicWord;echo $EnglishWord;
}}<?php
```

- **Performance Requirements**

The user may decide on a maximum acceptable error rate (or several rates

depending on the seriousness of the errors) and a minimum for the percentage of sentences considered grammatically acceptable in the raw output. (John and Laurent, 2016)

In this study the similarity percentage that is acceptable is shown in table 2.

Table 2: Acceptable rates categories

| Similarity percentages | Categories |
|---|---|
| 100% | Identical |
| More or equal 90% | Similar |
| 75% - less than 90% | acceptable |
| 50% - less than 75% | moderate |
| 40% - less than 50% | poor |
| Less than 40% | Not similar |

## CONCLUSION AND RESULTS

Some collected words in Arabic and English with the transliteration were collected to be used in the database DB as shown in table 3:

Table 3: work collections to be used in the DB

| Words (in Arabic | Transliteration | Translation (in English) | Similarity percentage |
|---|---|---|---|
| مقابر | Macabre | macabre | 100 % |
| شفط (يشفط) | Shaft | shaft | 100 % |
| ذات | That | that | 100 % |
| قانون | Canon | canon | 100 % |
| كوب | Cup | Cup | 100 % |
| سفري | Safari | Safari | 100% |
| غول | Ghoul | Ghoul | 100% |
| مسلم | Muslim | Muslim | 100% |
| سلطان | Sultan | Sultan | 100% |
| ليمون | Lemon | Lemon | 100% |
| أمين | Ameen | Ameen | 100% |
| قطن | Cotton | Cotton | 100% |
| جس | Guess | Guess | 100% |
| حريم | Harem | Harem | 100% |
| الظهير | Adhere | adhere | 100% |
| الجبر | Algebra | algebra | 100% |
| البنان | Banan (fingers) | banana | 100% |
| الجدول | Chedule | schedule | 85.71% |
| كاميرا | Camerh | Camera | 85.71% |
| البدوي | Bedoui | bedouin | 85.71% |
| الكحول | Alkohol | Alcohol | 85.71% |
| سبانخ | Spanikh | Spanish | 85.57% |
| صندل | Sandal | Sardal | 83.34% |
| أميرال | Amiral | Admiral | 83.34% |
| البرميل | Barmel | barrel | 83 % |
| سكر | Sugar | Sugar | 83 % |
| نادر | Nader | Nadir | 80% |
| حوش | Housh | House | 80% |
| كفر | Cofer | Cover | 80% |
| برك (يبرك) | Parak | Park | 80 % |
| الذيل | Thail | tail | 80 % |
| إذن | Ethen | then | 80 % |
| أقر | Agre | Agree | 80 % |
| عطر | Atter | Attar | 80 % |
| قناة | Cana | canal | 80 % |
| شاش | Shash | Sash | 80% |
| سماق | Sumaq | Sumac | 80% |
| دون | Doon (prep.) | Down | 75 % |
| أجل | Agel | Age | 75 % |
| أيّد | Aiid | Aid | 75 % |
| قول | Caol | Call | 75 % |
| زعفران | Safran | Saffron | 66.67% |
| قرنية | Carnia | Cornea | 66.67% |
| غزال | Ghazal | Gazelle | 66.67% |
| شيطان | Shitan | Satan | 66.67% |
| البق | Bag (insect) | Bug | 66 % |
| البوع | Elboo’ | Elbow | 66 % |
| قط | Ct | cat | 66 % |
| المتراس | Matrass | Mattress | 62.5 % |
| جرة | Jarra | Jar | 60 % |
| صلد | Sald | solid | 60 % |
| العنبر | Anbre | Amber | 60 % |
| كافور | Cafur | Camphor | 60% |
| وسط | Wasat | Waist | 60% |
| صفر | Zefer | Zero | 60% |
| علاوة | Allawah | Allowance | 57.7% |
| عتيق | Atiq | Antique | 57 % |
| كبّل | Kabbal (tie with rope) | Cable | 50 % |
| قابل | Qabbil (Adj.) | Able | 50 % |
| كف | Kaff (hand) | Cuff | 50 % |
| مزري | Muzry | misery | 50 % |
| عنق | Onk | Neck | 50 % |
| أزمة (ربو) | Azma | asthma | 50 % |
| زيركون | Zircon | Zirconium | 50% |
| عفريت | Afreet | Afrit | 50% |
| مومياء | Mumiya | Mummy | 50% |
| الخوارزمي | Algoarizmi | algorithm | 44.4 % |
| الإكسير | AlExire | Elixir | 42.85% |
| قنديل | Qandeel | Candle | 42.8% |
| شراب | Shrap | Syrup | 40% |
| قصر | Qaser | Castle | 40% |
| باقة | Baqah | Bouquet | 40% |
| مرآة | Merah | Mirror | 40% |
| مسجد | Masjed | Mosque | 33.4% |
| زرافة | Zrafa | Giraffe | 28.5% |
| مخزن | Makhazan | Magazine | 25% |
| كهف | Kahf | Cave | 25 % |
| لعق | Laak | Lick | 25% |
| خليج | Ghleaj | Gulf | 16.7% |
| قرمزي | Qurmosy | Crimson | 16.67% |
| جرثومة | Garthomah | Germ | 14.3% |
| البرقوق | Albrqooq | Apricot | 12.5% |
| قائد | Qaed | Guider | 0% |

## CONCLUSION

Arabic language is amongst the most highly spoken languages in the Muslim world. It is the native language of more than 28 countries. The revealance of Holy Book (Quran) made Arabic a sacred language for Muslims all over the world. It is equally important to native as well as nonnative speakers. Nonnative speakers, however face difficulty in understanding materials because of their lack of Arabic language skills. As well, the needs of spreading Arabic language is important to clear the Islamic and Arabic issues for the world, Current study was motivated to proposed approach that can use the advantage of the similarity between Arabic and English language to make the learning process easy for non-Arabic speakers to understand Arabic. The target approach adopted the functions

**(similar_text()),** (levenshtein())(Gosselin et al., 2011).

The main finding of the study is summarized by the following :

- The approach was target the Arabic words that have 40% and more similarity with the English language as shown in table 2.
- Table 4 illustrates the results percentage after implementation:

Table 4: Results of the study

| Similarity acceptance | No. words in percentage | Class of the word |
|---|---|---|
| 100% | 20.48 | Identical |
| More or equal 90% | 0 | Similar |
| 75% - less than90% | 28.98 | acceptable |
| 50% - less than75% | 30.12 | moderate |
| 40% - less than50% | 8.40 | poor |
| Less than 40% | 12.02 | Not similar |
| Total | 100% | |

- Table 4 indicates that 79.72% from the sample which translated and written in Latin letters can express the same words in English within the same meaning. Meanwhile, 8.40% cannot clearly describe the words which they were rejected by the proposed system for displaying, within the same reaction of the target approach 12.02% can not totally express the meaning of the words in Arabic language even of the similarity of the pronunciation between Arabic and English words ( see figure 9 ). Figure 10 illustrates the findings of the study.

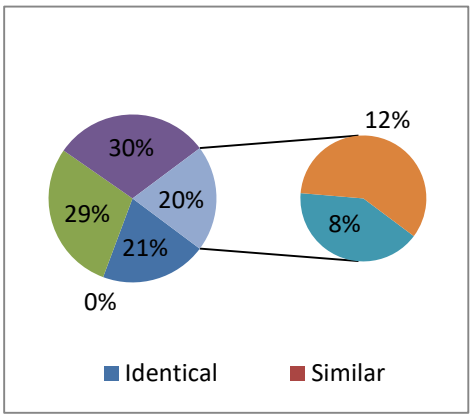


Figure 9: the percentage of accepted and unaccepted word regarding to the proposed approach.

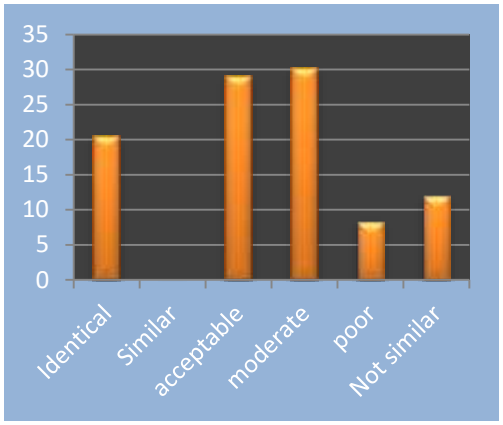


Figure 10: The findings of the study

ACKNOWLEDGMENT

We first wish to prostrate thanks to Allah the Almighty for having given us the courage and patience to finish this work. We then thank the dean of deanship of scientific research for having honored us by his coaching, his precious advice, his patience and his noble human values. A big thank you goes also to all the members of the University of Hail with whom we share much more than a place of work.

REFERENCES

[1] Aaron B., Philips, Violetta-Sforza (2012) "Arabic-to-English Example Based Machine Translation Using Context-Insensitive Morphological Analysis", 4th International conference on Arabic language Processing.

[2] Abu Naba'ih, A., Hussain, J., Al-omari, A., & Shdeifat, S. (2009). The Effect of Computer Assisted Language Learning in Teaching English Grammar on the Achievement of Secondary Students in Jordan. The International Arab Journal of Information Technology, 6, 431–439.

[3] Abufanas, O. (2013). Computer Aided Language Learning System for Arabic for Second Language Learners. International Journal of Social, Management, Economics and Business Engineering, 7(12), 1942–1946.

[4] Al-bazeli, A., Alreshidi, A. N. G., Naimie, Z., & Halili, S. H. (2014). Malaysian Students Learning The Arabic Language In Arab Schools In Malaysia : A Future Study. The Online Journal of New Horizons in Education, 4(4), 119–126.

[5] Baugh, A. C., & Cable, T. (2012). A History of the English Language (6th ed.). Longman.

[6] Bekele, T. A. (2010). Motivation and satisfaction in internet-supported learning environments: A review. Educational Technology and Society.

[7] Biller, Alia K. (2010) "Translation Recognition in Learners of Arabic", Inquiry journal, Volume 11, 2010.

[8] Bray, Allyson (2011). "Understanding the errors of arabic speaking ell's".

[9] Ciobanu, Alina Maria and Liviu P. Dinu (2014) "Automatic Detection of Cognates Using Orthographic Alignment".

[10] Dina, A., & Ciornei, S.-I. (2013). The Advantages and Disadvantages of Computer Assisted Language Learning and Teaching for Foreign Languages. Procedia - Social and Behavioral Sciences, 76, 248–252. doi:10.1016/j.sbspro.2013.04.107

[11] Ghalib, M., Hassan, H., Sahrir, M. S., Daud, N. M., Mohin, M., & Ayu, M. A. (2014). Developing Icollect Mobile (Android and Iphone-Based) Application for Arabic Language Teaching. Middle-East Journal of Scientific Research, 20, 21–27. doi:10.5829/idosi.mejsr.2014.20.lcl.203

[12] Goldberg, Yoav and Michael Elhadad, (2007) "Identification of Transliterated Foreign Words in Hebrew Script"

[13] Gosselin, D., Kokoska, D., & Easterbrook, R. (2011). PHP Programming With MYSQL (2nd ed.). Boston, MA: Course Technology, Cengage Learning.

[14] Guarda, M. (2012). Computer-mediated-communication and foreign language education. Journal of E-Learning and Knowledge Society, 8, 15–27.

[15] Hamida, N. (2012). Learning Styles and Strategies in Second Language Learning. Ragam Jurnal Pengembangan Humaniora, 12(1), 8–14.

[16] Hara, Y. (2011). The Use Of Loanwords In English Vocabulary Learning. BALL STATE UNIVERSITY, MUNCIE, INDIANA.

[17] Hussain, A. (2012). Textual Similarity. Bachelor thesis, Technical University of Denmark.

[18] Ibrahim, M. N., Bilmas, S. N., Babiker, A., & Idroas, M. (2013). A framework of an online self-based learning for teaching arabic as second language (TASL). In Proceedings of International Conference on Computational Intelligence, Modelling and Simulation (pp. 255–260). doi:10.1109/CIMSim.2013.48

[19] IJMES, International Journal of Middle East Studies, "IJMES Translation & Transliteration Guide". Retrieved 2016 December

[20] https://ijmes.chass.ncsu.edu/IJMES_Translation_and_Transliteration_Guide.htm

[21] Jethro, O. O., Grace, A. M., & Thomas, A. K. (2012). E-Learning and its effects on teaching and learning in a global age. International Journal of Academic Research in Business and Social Sciences, 2, 73–78.

[22] John Lehrberger, Laurent Bourbeau "Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation", John Benjamins Publishing, retrieved 2016 December

[23] Levitin, A. (2012). Introduction to the design & analysis of algorithms (3rd ed.). Boston: Pearson.

[24] Mahmoudi, E., Samad, A. B. A., & Razak, N. Z. B. A. (2012). Attitude and Students' Performance in Computer Assisted English Language Learning (CAELL) for Learning Vocabulary. Procedia - Social and Behavioral Sciences, 66, 489–498. doi:10.1016/j.sbspro.2012.11.293

[25] Malmasi, Shervin and Dras, Mark, (2015) "Cognate Identification using Machine Translation", Proceedings of the Australasian Language Technology Workshop (ALTA), 138-141. Sydney, Australia.

[26] Nachoua, H. (2012). Computer-Assisted Language Learning for Improving Students' Listening Skill. Procedia - Social and Behavioral Sciences, 69(Iceepsy), 1150–1159. doi:10.1016/j.sbspro.2012.12.045

[27] Otwinowska-Kasztelanic, A. (2012). Plurilingualism and Polish Teenage Learners of English. Lingvarvm Arena, 3, 37–52.

[28] Pudaruth, S., & Jannnoo, N. B. (2010). An E-Learning tool for the self-study of mathematics for the CPE examination. World Academy of Science, Engineering and Technology, 46, 427–432.

[29] S Norfaizah, & Muhammad. (2011). Perbualan harian : bahasa Arab, Melayu, Inggeris. Selangor, Malaysia: AR-Risalah.

[30] Saadane, Houda, Ouafa Benterki, Nasredine Semmar,Christian Fluhr. (2009) "Using Arabic Transliteration to Improve Word Alignment from French-Arabic Parallel Corpora"

[31] Sahrir, M. S., Yahaya, M. F., & Nasir, M. S. (2013). EZ-Arabic for Children: A Virtual Learning Resource Tool for Malaysian Primary Schools. Procedia - Social and Behavioral Sciences, 90(InCULT 2012), 396–404. doi:10.1016/j.sbspro.2013.07.108

[32] Sahrir, M. S., & Yusri, G. (2012). Online Vocabulary Games For Teaching And Learning Arabic. GEMA OnlineTM Journal of Language Studies, 12(September), 961–977.

[33] Salkind, N. J. (2010). Encyclopedia of Research Design. SAGE Publications.

[34] Schepens, J., Dijkstra, T., Grootjen, F., & van Heuven, W. J. B. (2013). Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. PLoS ONE, 8. doi:10.1371/journal.pone.0063006

[35] Shehab, Roaya et. al. "Arabic-Malay Cognates As A Computer Assisted Language Learning", 3rd International Conference on Advanced Computer Science Applications and Technologies (ACSAT 2014) 29-30 Dec 2014 Amman, Jordan

[36] Sherif, T., Kondrak, G., (2007) "Bootstrapping a stochastic transducer for Arabic-English transliteration extraction". In: Proc. of ACL.

[37] Soleimani, H., & Molla Esmaeili, M. (2012). The Effect of Explicit Instruction of Clustering New Words on Vocabulary Learning of Iranian Intermediate EFL Learners through Hyperlinks. International Journal of Applied Linguistics & English Literature, 1(2), 112–126. doi:10.7575/ijalel.v.1n.2p.112

[38] Stoimenova, S. T. (2010). Evaluation of Web Based Information Systems : Users ' Informing Criteria. Issues in Informing Science and Information Technology, 7.

[39] Suliman, I., Mukhtar, A., & Sahrir, M. S. (2013). Issues and Challenges of Using CALL in Learning Arabic : Considerations for Content Development. American International Journal of Social Science, 2(5), 76–88.

[40] Tatroe, K., MacIntyre, P., & Lerdorf, R. (2013). Programming PHP (3rd ed.). North, Sebastopol, CA: O'Reilly Media, Inc.

[41] Tavakoli, M., & Gerami, E. (2013). The Effect of Keyword and Pictorial Methods on EFL Learners ' Vocabulary Learning and Retention. Porta Linguarum, 19, 299–316.

[42] Webber, A. B. (2010). Modern Programming Languages: A Practical Introduction (2nd ed.). Franklin, Beedle & Associates Inc.

[43] Whiteman, M. (2012). Learning Spoken Arabic through Computer-Assisted Instruction: An integrated approach.

[44] Wong, L. L. C., & Nunan, D. (2011). The learning styles and strategies of effective language learners. System. doi:10.1016/j.system.2011.05.004