EUROPEAN ACADEMIC RESEARCH
Vol. II, Issue 12/ March 2015

Impact Factor: 3.1 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

Building an Automatic System to Construct a Thesaurus for Arabic Language Words

HASAN HUSSEIN RASHAIDEH
Prince Abdullah Bin Ghazi Faculty of IT
University of Al-Balqa' Applied, Salt
Jordan
IMAN HUSSEIN AL-QINANI
Education College
University of Al-Mustansiriyah, Baghdad
Iraq
SALEH AL OQEILI
University of Al-Balqa' Applied, Salt
Jordan

Abstract:

*Constructing an automatic thesaurus for Arabic is a very challenging and difficult task due to several reasons. In this paper we proposed new method to constructing an automatic system thesaurus for Arabic language words. In order to achieve this objective, we have selected and tested a new approach for building an automatic system to construct a thesaurus for Arabic language words. We have applied the suggested approach on Al Wasit database (**Al Wasit dictionary**), where the most frequent words and the meaningless ones are cancelled.*

This research has relied upon the statistical instead of the linguistics methods in building the system. In addition, it have offered several alternatives to be used in the case of improper results of the program, such as re-ranking the results, adding unavailable word in dictionary or supporting the word definition which is found in the dictionary through entering an article relevant to the entered topic of the word.

The suggested method give promised results and it will used different strategies from other methods. Time Complexity for this algorithm is equal to O (n2).

Key words: Thesaurus, Arabic language, Al Wasit dictionary, Stopword, Rank.

1. Introduction

Technology has touched every aspect of our lives. Literally, it changed the way we live and conduct our daily activities. People rely heavily on technology in their homes, school, and work place. Reading books has become through digital media and looking up words from a dictionary or thesaurus has become electronic as well (Al-Qabbany, *et al.,* 2009; Turban, *et al.,* 2002).

The study of the constructing a thesaurus for Arabic language words gaining importance because it tries to keep pace with technology of NLP (Al-Qabbany, et al., 2009; Turban, et al., 2002). Nowadays, people do not have time to spend on looking up a word in a dictionary or thesaurus. Thus, they rely on technology to maximum extent (Turban, et al., 2002). Most of them carry pocket size electronic dictionaries or thesaurus which can fulfill their need instantly (Turban, et al., 2002).

Constructing an automatic thesaurus for Arabic in particular is a very challenging and difficult task due to several reasons: Thesaurus for Arabic language is very rare whereas most of the studies, research and thesaurus construction approaches are conducted and designed for the English language (Zaidi, et al., 2005). As a result, before examining a thesaurus constructing approach, this research will deal with the Arabic language limitation's such as lack of database for words of similar meaning and how to apply a construction method such as Soergel 1974 (William, et al., 1992) or Grefenstette's syntactical contexts (Senellart, et al., 2008) and others which have been designed for building a thesaurus for English and how to apply them on the Arabic language.

Arabic language is not receiving or having sufficient and proper sources from both hand written and electronic

references. As a result, constructing a thesaurus is a challenging task. The study is restricted to the spoken Arabic or the easy version of it since most of the inside and outside of Arabs do not know the meaning of old Arabic literature vocabulary and philology (Khafajeh, et al., 2010). Arabic language lacks the databases that contain a large volume of Arabic vocabulary. As a result, gathering and categorizing words similar in meaning is a very challenging task.

Many researchers will be conducted about the similarities between the English and the Arabic terms in order to reveal the closest approach of building a thesaurus and its application in the Arabic language. At the same time, a comparison will be conducted between the manual approach (Senellart, et al., 2008; William, et al., 1992) and the automatic one (Al-Qabbany, et al., 2009; Hsinchun, et al., 1995; Khafajeh, et al., 2010; Senellart, et al., 2008; William, et al., 1992) in order to declare which one of them is more appropriate to construct a thesaurus for Arabic language words. However, The study of this subject has few limitations such as Arabic letters being different from other language letters (Khafajeh, et al., 2010) because Arabic letters are subject to: Alkasrah, Dammah, Fathah and sukun according to these mark the letter is pronounced (Khafajeh, et al., 2010; Zaidi, et al., 2005): **[نْ، نَ ، نُ ، ن]**

(Zaidi, *et al.,* 2005) presented a method to describe a web–based multilingual tool for Arabic information retrieval based on ontology in the legal domain. They started with the manual construction of the ontology and its editing via protégé 2000. The study used (UN: United Nations) Arabic and query expansion for Arabic documents. Query expansion is achieved through using a semantic word thesaurus – Word net; Results showed that there is a significant improvement in the recall and the precision.

(AL-Qabbany, *et al.*, 2009) proposed an improvement to the similarity thesaurus construction method used for query expansion in information retrieval as "MEAN" method; the proposed improvement of about 3.3% over the SUM method. Results "MEAN" method more accurate than the "SUM" method and it can discover and eliminate the outlier. Source data used the France Press Agency news as the document collection. Number of documents (208,596), number of terms (435,846), number of terms occurrences (30,415,222), number of processed terms (248,311), average number of words per document (69.78).They have choose twenty general topics to use for the evaluation process.

(Khafajeh, et al., 2010) discussed a major problem of modern Information Retrieval (IR) systems, which is the vocabulary problem that concerns the discrepancies between terms used for describing documents and the terms used by the researchers to describe their information need. Using a thesaurus is one way to overcome vocabulary problem. (242) Arabic texts were used and 59 Arabic queries. All of it involved computer science and information system. The main objective of the paper is to design and build automatic Arabic thesaurus. They used term-term similarity and association techniques for every field and domain. Results showed that the association thesaurus improved the recall and precision over the similarity thesaurus.

2. Proposed Methodology

The work of this research goes through several steps in order to achieve the objectives of the study which aims at constructing an Arabic automated thesaurus. The following subsections illustrate each step in building an automatic system to construct a thesaurus for Arabic language words. The suggested algorithm showed in Figure 1.

2.1 Obtaining Al Wasit Data

In this step, we aim at inserting Al Wasit dictionary data word file into excel file with two columns: the first column contains the word and the second column contains the definition of the word. The two columns are in Arabic language as the examples showed in Table1.

Table 1: Al Wasit Data in Excel File

Word	Definition
أسيا	قارة في الكرة الأرضية
السكري	البول السكري مرض يظهر فيه سكر العنب في البول نتيجة لأسباب متعددة أهمها نقص هرمون الأنسولين الذي (ينظم احتراق هذا السكر في خلايا الجسم (مج
الشهر	جزء من اثني عشر جزءا من السنة (الشمسية و القمرية) و يقدر في السنة القمرية بدورة القمر حول الأرض و يسمى الشهر القمري أو يقدر بجزء من اثني عشر جزءا من السنة الشمسية و يسمى الشهر الشمسي (ج) أشهر و شهور و الأشهر الحرم الأشهر التي كانوا يحرمون فيها القتال و هي أربعة ثلاثة منها متوالية و هي ذو القعدة و ذو الحجة و المحرم و واحد فرد و هو رجب و نظام رسمي لتوثيق العقود و نحوها و إعلانها (محدثة) و (مصلحة الشهر) إدارة حكومية قائمة على توثيق العقود و نحوها (محدثة
المدرسة	مكان الدرس و التعليم و جماعة من الفلاسفة أو المفكرين أو الباحثين تعتنق مذهبا معينا أو تقول برأي مشترك (مج) و يقال هو من مدرسة فلان على رأيه و مذهبه (ج) مدارس

2.2 Create Table in Oracle Database

The second step to create database table in Oracle SQL based on Al Wasit dictionary excel file taken from the previous step that has been inserted in (Data table). At the end of this step, we have Al Wasit data in Oracle database. as shown in the examples in Figure 2.

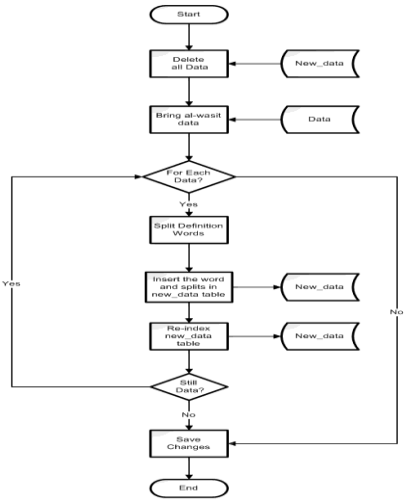


Figure 1: The suggested algorithm.

2.3 Data Analysis

Third step is to separate the word and its definition in simple table automatically, and to insert the process data in it (New Data table). The separation has been made based on space character between definition words which makes (609551 Sub Data) as shown in the example in Table 2.

Table 2: New_Data Table

Word	Sub Data
آسيا	قارة
آسيا	في
آسيا	الكرة
آسيا	الأرضية
المدرسة	مكان
المدرسة	الدرس
المدرسة	و التعليم
المدرسة	و جماعة
المدرسة	من
المدرسة	الفلاسفة
المدرسة	أو
….	…..

Figure 2: Al Wasit Data in Oracle Database

2.4 Remove Remarks

Now, we have to remove all the unwanted characters and replace them with null value in Sub Data. For example: (",","!","","؟","(",")",";","\","/",":","&","<",">","^","9","8","0","1","2","3","4","5","6","7",…etc).

2.5 Remove Stopword and Maximum Repetitions Words

Stopword means the words which are very frequent and they are considered as weak to be distinguished, we cannot distinguish the content of a text depending on them. These words are removed and not indexed in order to save both space and time (Khafajeh, et al., 2010). Here, Stopword includes: (demonstrative pronouns, relative pronouns, Arabic letters, names and verbs, interjection verbs, conditional pronouns, interrogative pronouns, etc.). It is shown in the examples given in Table 3.

At this step, we create two tables: the first is named (**Stopword Table**) which contains most of the popular stopword in Arabic and it contains about (14,000 words), they were obtained from the World Wide Web. Removing and Reducing thousands of stopword will speed up the search in the algorithm.

The second table is named (**Maximum Repetitions Words Table**) which contains the most frequent words in (sub data column) in Table 2 as shown in Table 4. The words repeated more than 200 times in sub data column in Table2 have been included in Maximum Repetition Words Table.

2.6 Obtaining word definition

If there is no definition or the results are useless to the user, the user can bring the definition accompanied with the article relevant to the topic. Furthermore, if there are newer words which are not available in the dictionary, such as computer or globalization, the user can insert such words through the article.

Table 3: (A) Stopword Table. (B) Maximum Repetitions Words Table. (C) Statistics for Max Repetitions Words Table.

Word	Word	Word	Word
آخر	إذ	ثم	لم
أبدا	إذا	جدا	لما
أحد	إطلاقا	جميعا	لمدة
أحيانا	إلا	حاشا	لن
أخرى	إلى	حاليا	لنا
أخيرا	إلي	حتى	له
أشياء	إليك	حول	لها
ألا	إليكم	حولك	لهذا
أما	إليكما	حولكم	لهم
أمام	إليكن	حولكن	لهما
أمامك	إلينا	حولنا	لهن
أمامكم	إليه	حولهم	لو

A

Word	Repetition
و	51131
من	14288
في	13714
ج	8529
يقال	8465
أو	6282
ما	5024
على	4818
به	3828
ويقال	3107
لا	2806

B

Process	Result
Words Count	205
Max. Repetition	51131
Min. Repetition	202
Avg. Repetition	1151
Sum. Repetition	235961

C

2.7 Search Algorithm

For searching synonyms of any given word, search algorithm look about the input word on Word column for Table 2 (New_Data Table) and when it found it return the corresponding word in the sub data column, otherwise the new word will insert in the table with some identification about it (this can happen for new or updates words. The result table will be rearrange to remove all the redundant words.

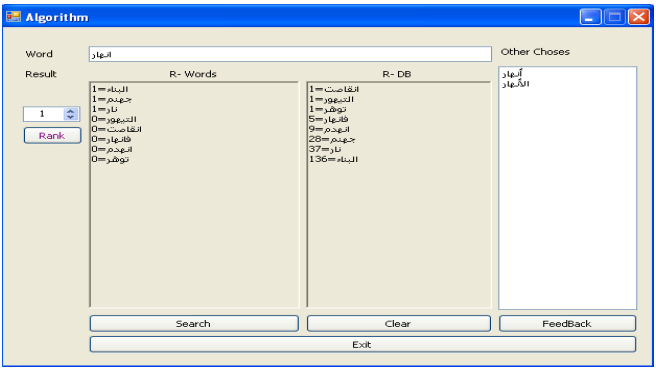
The system gives the user some other choices such as ranking the result and the rank mark insert in the rank table. The user can also give feedback, new word and its definition, or both in algorithm. The user can Re-Rank the result and change the algorithm result if there is an undesired result.

2. 7.1 Other Choices Result Process

The problem of writing the letter in more than one form (it is normal in Arabic language) makes it possible for the user to be confused. As a result, this leads to changing the form of the word, which in turn leads to changing the meaning of the word. The letters that can be written in different forms and consequently cause confusion include the letter (ا) which can be written in the forms (أ، آ، إ، ا), the letter (و) which can be written in the forms (و، ؤ), the letter (ت) which can be written in the

forms (ت، ة، ه) and the letter (ى) which can be written in the forms (ئ، ى، ي). This procedure aims at processing the forms of the following letters (ا، ت، ؤ، ى) as well as defining the word with article (ال) or subtracting it from the word. Then, all the possible forms of the letter in the inserted word are created as shown in the example given in Figure 3.

Figure 3: Example for Other Choices Result Process



2. 7.2 Feedback Process

The goal of this procedure is to clarify the process whether the user has entered a certain synonym into the feedback domain. Several updates shall occur through this entry, such as accrediting the new word in word_ feed table as well as the Ranking and approving of the synonym within the results.

2. 7.3 Rank up/down Button Process

Rank, which means that synonymous produced by the system relative of the original word (synonyms arrangements which the system produces are closer to the meaning.). High rank means closer synonymous to original word, rank start numbering from 0 to infinite. Using this procedure, processing occurs if the user presses the Rank Button. By doing so, the user changes the rank of the word; thus, the rank will update. In addition to that, update of the word_ feed table occurs too, as the example shows in the Figures 9.

2.8 Algorithm Application Mathematically

Algorithm applicable in the form of mathematical equations:

$$
\begin{aligned}
&Assumptions\\
&X = SearchWord \rightarrow\rightarrow Find \therefore f(X)\\
&O(X) = OtherChoicesMetrix\\
&F(X) = FeedbackMetrix\\
&DB(X) = LeftSideMatrix \bigcup RightSideMatrix\\
&R(X) = F(X) \bigcup DB(X)\\
&R(X) = Distinct(R(X))\\
&ForAllElementInMetrix :\rightarrow R(X)\\
&Calculate\\
&NR(X) = (R(X1), R(X2), ..., R(Xn))\\
&Rank\ \mathrm{Re}\ sultMatrix = R(X) \bigcap NR(X)
\end{aligned}
$$

Where: X is the input word, **O(X)** is the other choices for the input word such as (إيمان ,أيمان,ايمان: ايمان)}, **F(X)** is the feedback, which is the closest synonym to the user input word almost entered by the language expert, **DB** is the words collection that we got from New_Data Table, **R(X)**: is the words collection that we got from **F(X) and DB, Distinct (R(X))** delete the repeated words form **R(X).**

NR(X): Find **R(X)** for each synonym result from **(R(X)** for input word)

Rank Result Matrix: find rank of the synonyms by intersect **(R(X)** for input word) with **NR(X)**.

2.9 Example for search Algorithm

Table 4: Example for Word Processing [المدرسة]

Word Processing	The Original Definition	After Process Definition	Right Sides Words Not Include in the Definition
المدرسة	مكان الدرس و التعليم و جماعة من الفلاسفة أو المفكرين أو الباحثين تعتنق مذهبا معينا أو تقول برأي مشترك (مج) و يقال هو من مدرسة فلان على رأيه و مذهبه (ج) مدارس	الدرس	الصف
		التعليم	الفرقة
		جماعة	الفصل
		الفلاسفة	الإمام
		المفكرين	الناظر
		الباحثين	الدرس
		تعتنق	
		مذهبا	
		معينا	
		مشترك	
		مدرسة	
		رأيه	
		مذهبه	
		مدارس	

Table 5: Create R(X) for Word [المدرسة]

X	O(X)	F(X)	DB(X)		R(X)	Distinct(R(X))
	المدرس	ثقافة	LeftSideMatrix	RightSideMatrix	الدرس	الدرس
	مدرسة	ثانوية	الدرس	الصف	التعليم	التعليم
			التعليم	الفرقة	جماعة	جماعة
			جماعة	الفصل	الفلاسفة	الفلاسفة
			الفلاسفة	الإمام	المفكرين	المفكرين
			المفكرين	الناظر	الباحثين	الباحثين
			الباحثين	الدرس	تعتنق	تعتنق
			تعتنق		مذهبا	مذهبا
			مذهبا		معينا	معينا
			معينا		مشترك	مشترك
المدرسة			مشترك		مدرسة	مدرسة
			مدرسة		رأيه	رأيه
			رأيه		مذهبه	مذهبه
			مذهبه		مدارس	مدارس
			مدارس		الدرس	الصف
					الصف	الفرقة
					الفرقة	الفصل
					الفصل	الإمام
					الإمام	ثقافة
					ثقافة	ثانوية
					ثانوية	الناظر
					الناظر	0

Table 6: NR(X) for Word [المدرسة]

(R(X))	NR(X)													
الدرس	الجرب	العلم	ثلب	المدرس	يدرس	أدراس	البالي	المقدار	وقت	دروس	درسان	المدرسة	البرنامج	...
التعليم	التلوي	الفرقة	المدرسة	الأساس	الطالب	المعلم	المنهاج	الثنيء	/	/	/	/	/	/
جماعة	الخيط	الجملة	الجيش	السلك	اختزره	المدرس	الأخوش	الحصوية	الحضر	الجلسة	النملة	المدرسة	التسعار	...
الفلاسفة	الجدل	الجمال	الملة	الحب	الإبداع	المدرسة	تفلسف	الفطرة	العدالة	الجسم	التكاكون	السلبية	/	/
المفكرين	الخلوصية	المدرسة	/	/	/	/	/	/	/	/	/	/	/	/
الباحثين	المدرسة	/	/	/	/	/	/	/	/	/	/	/	/	/
تعتنق	المدرسة	/	/	/	/	/	/	/	/	/	/	/	/	/
مذهبا	أسخى	الصاحب	المدرسة	المذهب	ذهب	مجمج	/	/	/	/	/	/	/	/
معينا	وجه	الجلسي	الجهاز	الدولة	الكمبيالة	المدرسة	/	/	/	/	/	/	/	/
مشترك	الأسرة	التشكيك	التعاون	التركة	المدرسة	المشترك	النوع	/	/	/	/	/	/	/
مدرسة	تثقف	المدرسة	/	/	/	/	/	/	/	/	/	/	/	/
رأيه	صمم	فد	أعصله	أفد	استعمله	المستشار	المدرسة	الخلل	المتزمت	القعمان	خدع	صمم	طمن	...
مذهبه	السلوك	حرض	أحمد	الأسلوب	تحليل	المدرسة	الداخلة	/	/	/	/	/	/	/
مدارس	الحصاصلة	المدرس	المدرسة	/	/	/	/	/	/	/	/	/	/	/
الصف	الفصل	المدرسة	صفا	اتباع	الأسكوب	الأسلوب	المنمماك	المصطفون	تراصف	يقاتلون	صفوف	المستقيم	الرزدق	...
الفرقة	البين	التعليم	الشحنة	الجيش	الفريق	المدرسة	عدد	الألعاب	المطافئ	الفئة	الشحنة	التمثيل	الاختراق	...
الفصل	والشتاء	الربيع	الخطاب	السنة	التمسية	القوامة	الكتاب	الصف	المدرسة	أربعة	التمثيلية	والصيف	فصول	...
الإمام	الحذيلي	اتبع	الإمامة	التلميذ	القرآن	القدر	الإمة	المدرسة	حفظ	تنفع	أئمة	الصلاة	للمسلمين	...
ثقافة	ثقف	/	/	/	/	/	/	/	/	/	/	/	/	/
ثانوية	/	/	/	/	/	/	/	/	/	/	/	/	/	/
الناظر	الناظرة	إدارة	القبال	المجس	الوزير	نواظر	والمين	نظر	المعارف	نظارة	المدرسة	الرادار	البولة	...

After Rank Result Matrix, the algorithm will show the result in a descending order depending on Rank Result Matrix as shown in Table 9:

Table 7: f(X) for Word [المدرسة]

Distinct(R(X))	Rank Result	Discuss the Rank Results
الفرقة	3	Intersect by التعليم + 2 (up button Rank)
الصف	2	Intersect by الفصل+ (up button Rank)
الفصل	2	Intersect by الصف + (up button Rank)
الإمام	1	(up button Rank)
ثقافة	1	(Feedback)
تقوية	1	(Feedback)
التعليم	1	Intersect by الفرقة
الباحثين	0	From Original Definition
رأيه	0	From Original Definition
معنا	0	From Original Definition
متشترك	0	From Original Definition
مذهبه	0	From Original Definition
مذهبا	0	From Original Definition
مدرسة	0	From Original Definition
مدارس	0	From Original Definition
جماعة	0	From Original Definition
تعلّق	0	From Original Definition
الفلاسفة	0	From Original Definition
المفكرين	0	From Original Definition
الناظر	0	From Original Definition
الدرس	0	From Original Definition

3. Experimental Results

In order to assess building a system for the construction thesaurus for Arabic language words, we have developed an assessment tool that is a questionnaire. The developed questionnaire has been distributed over to a group of experts who are college professors in Arabic language along with other faculty members and PhD. students in several Jordanian and Arab Universities. This sample consists of (41) participants.

Set of Arabic words has been randomly chosen from "Mujam Al Wasit." These words are arranged in the assessment questionnaire in order to examine the extent of matching of synonymous meaning which the system points out. Thus, it would be possible to look in arrangement of relevant words, so as to know the synonyms arrangements which the system produces are closer to the meaning.

3.1 Validity of Study's Tool

Validity of the study's tool has been obtained through submitting it to a panel of arbitrators who are experts in the field and are specialized in the Arabic language. an example for System Results for Words of the Questionnaire that it distributed over to a group of experts to assessment shown in the Table 8.

Table 8: example for System Results for Words of the Questionnaire

ت	الكلمة المدخلة	نتائج البرنامج لكلمات الاستبانة (ترتيب الكلمات حسب ظهورها في البرنامج)	الرقم يمثل الترتيب حسب القرب للكلمةالمدخلة
1	الشهد	عسل	3
		النحل	2
		الرصفة, سدى, دام	1
		الطرم, يمصر, شهدة, شهاد, القطمة, تسمعه, المزج	0

3.2 Procedures for Assessment Questionnaire Analysis

For the purpose of SPSS analysis of the data collected from the members of study's sample, responses are given the following scale (excellent, very good, good, satisfactory, weak). Each selection is given the rate as it is shown in the Table 9.

Means of the responses of the participants are obtained through conducted SPSS program for the purpose of results discussion, obtained scores are distributed on three levels as it is shown in the Table 10:

Table 9: The Scale and the Rate

NO.	Scale	Rate
1	**Excellent**	**5**
2	**Very good**	**4**
3	**Good**	**3**
4	**Satisfactory**	**2**
5	**Weak**	**1**

Table 10: Scores and Assessment Level

NO. Level	Scale	Assessment Level
Level 1	1 to 2. 33	Low
Level 2	2.34 to 3.66	Medium
Level 3	Exceeds 3. 66	High

Range level is computed through the following computation steps:

- The highest rate (Excellent) - lowest rate (weak) = (5-1) = 4
- The division of the product on the number of levels which the study has determined namely three levels: (Low, Medium, and High).
  4 / 3 = 1.33 the product is the level's length then addition the product to the lowest rate that is: 1+1.33= 2.33 thus, first level (low level) is (1 to 2.33),
  2.3+ 1.33 = 3.66 thus, second level (medium level) is (2.34 to 3.66).
  3.66 + 1.33 = 4.99 Thus, the rate exceeds 3.66 is high which is the third level.

3.3 Assessment Analysis

Means of the responses of the study's participants are calculated through statistical analysis. The results showed that most words in the assessment paper have means higher than (3.66) which they fill in the third level (High). This indicates that the order of synonyms which the system produces is close to a higher rate as showed in table 13.

Means of the responses of the study's participants are calculated through SPSS analysis as in the example of [الشهد]:

Table 11: SPSS analysis for word[الشهد]

Word	Excellent 5	Very good 4	Good 3	Satisfactory 2	Weak 1	Mean	Word Percentage
الشهد	***35***	***3***	***2***	0	***1***	4.7317	94.634

Accumulative Multiplication Scale = 35*5 + 3*4 + 2*3 + 0*2 + 1*1 = 194
Mean = 194 / 41 = 4.7317, Word Percentage = (4.7317/5)*100= 94.634

Table 12: Assessment Analysis for Words

No.	Word input to system	Mean	Assessment Level	Word Percentage
1	الشهد	4.7317	High	94.634
2	الرحراح	3.5122	Medium	70.244
3	الجشاء	4.3415	High	86.83
4	المجسم	3.9512	High	79.024
5	علماء	3.9756	High	79.512
6	شحب	3.2683	Medium	65.366
7	الفرات	4.4146	High	88.292
8	الجمان	3.3415	Medium	66.83
9	إختزه	3.4878	Medium	69.756
10	رجبي	4.6341	High	92.682
11	الرجولة	4.2439	High	84.878
12	الأرجوحة	3.6585	Medium	73.17
13	المدرسة	3.7073	High	74.146
14	الكوكب	4.6829	High	93.658
15	الكوثر	4.8049	High	96.098
16	اللجين	4.5610	High	91.22
17	الأردن	3.6829	High	73.658
18	الرحاق	4.4390	High	88.78
19	الطيان	4.3415	High	86.83
20	هيثاء	2.7805	Medium	55.61
21	السكري	4.7805	High	95.61
22	البلكرياس	4.5610	High	91.22
23	الدياج	4.3415	High	86.83
24	دمشق	4.5854	High	91.708
25	المجداف	3.4390	Medium	68.78
26	الوجز	3.5854	Medium	71.708
27	الوجنة	4.5366	High	90.732
28	التوحيد	4.7805	High	95.61
29	الصابون	4.5122	High	90.244
30	الضابطة	4.3415	High	86.83
Total Word Percentage				**2480.49**

The word percentage is given the ratio of acceptance for each word in the questionnaire by a group of experts.

Finally, the Total Word Percentage (T.W.P.) = 2480.49, the average percentage (AVG.) of whole questionnaire can be calculated using:

AVG. = (T.W.P. / No of Word) = (2480.49/30) = (82.683) so the accuracy of the questionnaire will be (83%)

Based upon assessment process which was conducted by the study's sample (Instructors and Arabic language teachers), there is a match among the meaning of the synonyms that the system produces. Furthermore, there are words which have relationships with synonyms that the system produces.

According to the word orders produced as synonyms. Thus, their order reveals that they are closer to meaning, thus it can be relied upon this system and consider it as a thesaurus for the Arabic language. However, it is important to conduct some improvement and in accordance with the system user's perspective.

3.4 Comparison

Compare this study with other papers for Building an Automatic System to Construct a Thesaurus for Arabic Language Words:

1. Other studies specialized in just one area as a legal, medical, while this study general.

2- This study searched for synonyms for the word using dictionary that Arabic words database in the event was word modern such as globalization (العولمة), computer(حاسوب) and (ديمقراطية) will add an article about this words to extract synonyms.

3. Using dictionary and articles as well as enter of expert user the near synonym to the word (feedback).

4- We have not used diacritized texts; however, we have attempted to remedy the forms of the letters that cause the user to make an error or that give the word a different meaning such as the letters (ا, ت, ؤ, ى).

4. Conclusions

We have to go to the original source of work; i.e. a database of Arabic words taken from Al Wasit Dictionary. This database is processed in several steps to constitute an automated thesaurus. As a part of the research, statistical intersection approach is used to obtain the very close synonymy to the original word.

A questionnaire is developed and analyzed in order to assess the words provided to reveal the extent of fitness or match the synonymous words produced by the system to the meaning or close to it. Results reveal that the **rate of assessment** is (83%), which indicates the success of the current study. **Time Complexity** for this algorithm is equal to $O(n^2)$ for both data result and database result because we use two loops in the maximum for building the system. **There are many advantages** of the system used in this study, namely:

first, Feedback techniques are simple and give good results. The second, changed ranking for the result return and re-order are completed easily. Third, Fast algorithm to process and get the result within less than a second for each search operation. Fourth, optimal memory usage no redundant data because database concept applies as database of the dictionary and other data. **Contributions of this study:** first,the study is important because it discusses an important subject that is constructing a thesaurus for Arabic language words. There is no accredited thesaurus for Arabic, even in its traditional form as a book or printed dictionary. Those unaccredited thesauruses do not include all the Arabic words that Arab speakers use. Second, the significance of this study can be employed in many Linguistic applications as Information Retrieval.

The disadvantages of the system used in this study are: first, Data relationship reflects intersected results: if data contains good definitions, the result will be perfect and near 100%; and if not, the results will be poor. Second, the algorithm needs time for the first initialization, only the first time to calculate the relationships. The third, using not full and partial diacritized texts.

REFERENCES

Al-Qabbany, A., Al-Salman, A., & Almuhareb, A.(2009). An Automatic Construction of Arabic Similarity Thesaurus. 3rd International Conference on Arabic Language Processing (CITALA2009), PP: 31-36, Rabat, Morocco.

Holy Quran. (n.d.).

Hsinchun, C., Yim, T., & Fye, D.(1995). *Automatic Thesaurus Generation for an Electronic Community System*. Journal of The American Society for Information science, 46:175-193.

Khafajeh, H., Yousef, N., & Kanaan, G. (2010). *Automatic Query Expansion for Arabic Text Retrieval Based on Association and Similarity Thesaurus*. European, Meditrranean & Middle Eastern Conference on Information Systems (EMCIS 2010) Global Information Systems Challenges in Management. PP: 1-16, Leroyal Meridien Abu-Dhabi, UAE.

Senellart, P. & Blondel, V.(2008).*Automatic Discovery of Similar Word. in: Survey of Text Mining ii* (B. M.W., Ed.). Springer-Verlag, pages: 25-44, London.

Turban, E. F., Lee, J., King, D., Warkentin, M., & Chung, H.(2002). *Electronic Commerce* (2 Ed.). A Managerial Perspective. Upper Saddle River, NJ: Prentice Hall. Hardcover: 914 pages, London, UK.

William, B. F., & Yates, R. B. (1992). *Information Retrieval: Data Structures and Algorithms* (1 Ed.). Prentice Hall, Hardcover: 473 pages.

Zaidi, S., Laskri, M., & Bechkoum, K. (2005). *A Cross-language Information Retrieval Based on an Arabic Ontology in the Legal Domain*. The International Conference on Signal-Image Technology & Internet–Based Systems (SITIS05), PP: 86-91, Morocco.

Hasan Hussein Rashaideh. University of Al-Balqa' Applied**,** Prince Abdullah bin Ghazi Faculty of Science and Information Technology, Salt, Jordan. He received MSc, and PhD in Computer Science. He has many published papers in the computer science field. His research interests are in image processing, biomedical, Artificial Intelligence, and Natural Language Processing. He Assistant Professor in Prince Abdullah bin Ghazi Faculty of Science and Information Technology at the University of Al-Balqa' Applied, Salt, Jordan. E-mail: rashaideh@bau.edu.jo

Iman AL-Qinani. University of Al-Mustansiriyah, Education College, Baghdad, IRAQ. She received BSc in computer science from AL-Mustansiriyah University, and MSc in computer science from Al-Balqa' Applied University, worked in university. She has many published papers in the computer science field. Her research interests are in image processing, biomedical, and Artificial Intelligence, She's Associate Teacher in Computer Science at the University of Mustansiriyah – Baghdad, IRAQ. E-mail: mcs_0000@yahoo.com

للفهم اكثر هنا الشرح بالعربي
مقارنه هذه الدراسة مع بقية الدرسات في بناء نظام لمرادفات اللغة العربية :
١- بقية الدراسات اختصت في مجال واحد فقط كان يكون قانوني ، طبي بينما هذه الدراسة كانت عامه
2- البحث عن مرادفات الكلمة باستخدام قاعدة بيانات القاموس الشامل لاغلب الكلمات العربية وفي حال كانت الكلمة حديثة مثل العولمة او الديمقراطية يتم اضافة مقال عن هذه الكلمة لاستخراج المرادفات لها.
3- تم استخدام القاموس والمقالات اضافه الى ادخال المسخدم الخبير المرادفات القريبة الى الكلمة .
4- نحن لم نستخدم الكلمات المشكلة تشكيل تام او تشكيل جزئي وانما عالجنا فقط اشكال الاحرف.

في الخلاصة كتبت :
الطريقة المستخدمة في هذه الدراسة تم البحث في قاعدة بيانات الكلمات العربية من خلال البحث في حقل كلمة وحقل تعريف الكلمة في جدول قاعدة البيانات عن الكلمة المدخله وذلك لضمان الحصول على المزيد من المرادفات، ومن ثم العمل على ترتيب المرادفات من خلال عملية التقاطع .