



information age. In this paper, the work of Al-Khalil was extended into the mathematical set theory and an equation was derived to predict the cardinal of the Arabic root words from given set of characters.

1.1 Introduction:

The accumulation of information increases every year, as researchers Hal Varian and Peter Lyman measured the total production of all information channels in the world for two different years, 2000 and 2003 and found 66% increase in the size of information in three years period [3]. Merriam-Webster Dictionary reports new words each year and the searchable text database, which linguists call it corpus, includes more than 70 million words. New words are coined in other languages each year.

The work of Al-Khalil for defining sets of root words with characters lengths of 2, 3, 4, and 5 is an efficient linguistic model for the generation of new words. Comparing Arabic characters patterns with existing words character patterns it was found that the most used patterns where in the 2 and 3 characters range. Considering the pattern of characters as encoding algorithm for words, it can be seen that the most redundantly used words were assigned patterns with the fewer characters lengths. This indicates that Arabic language is naturally efficient language for compressing data, as the derived words can be reverted to root words for reduction in characters lengths through the use of programmable morphological relations between root words and derived words. This was discussed in section 3.1. Al-Khalil work demonstrates that Arabic language is a natural example for modern Information Theory [19, 20] guide lines for data compression to save electronic storage and transmission bandwidth. Even direct translations into Arabic with the same eloquence for a given script from languages, such French or English languages, will reduce the storage size to the range of 30-40% that is before implementing compression of the same data in the Arabic language.

Al-Khalil lemmatization work on the Arabic language and his mathematical approach for his permutation of Arabic characters into words, made it



compatible with Algorithmic Information theory and set theory axioms, an example is the use of set theory in relational database [21, 23].

2.1 Patterns of Arabic Characters

A written word is a visual pattern of graphemes or characters. In this section Arabic visual pattern generation algorithm will be discussed. And derivation of universal equation for Al-khalil permuted sets of patterned root words will be introduced.

Figure 1, shows two sets. Set **B**, which is a set of characters with cardinal **b**. And set **G**, which is a set of spaces where characters can be placed cursively. Set **G** has a cardinal **g**. A cardinal is a natural number that is bigger than zero, which gives the set size or the count of elements in a given set.

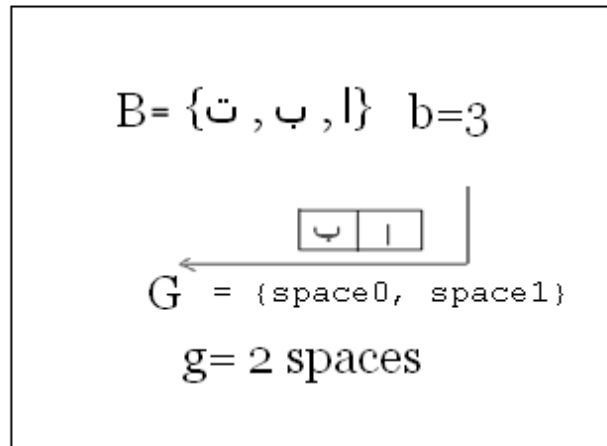


Figure 1: A visual pattern of a word is generated by spaces where the characters are placed, and the possible number of characters that can be used in each space by selecting one character at a time. **B** is the set of characters with a cardinal $b=3$. **G** is the set of spaces with cardinal $g=2$. **A** is a set of patterns for the elements in set **B**, that occurs due to the combinations of set **G** with the set **B** elements.



The combinations of set **G** elements with set **B** elements will generate set **A** as shown below.

$$A = \{\{ا\}, \{ا, ب\}, \{ا, ت\}, \{ا, ب, ت\}, \{ب\}, \{ب, ت\}, \{ت\}, \{ب, ت, ا\}, \{ت, ب, ا\}, \{ا, ب, ت, ا\}\}$$

The cardinal for set **A** is given [22] by the equation (1) below.

$$a = b^g \dots \dots (1)$$

To generate a set of patterns that resemble words by using high level computer language code, nested loops can be used. The nested loop depth is equivalent to **g**, and the upper limit for each nested loop is **b**. C++ code is given below for the nested loops.

```
for(int i=0; i<b; i++)
    for(int j=0; i<b; j++)
        .....
    Conditions;
```

The nested condition inside the last inner loop, were mainly used to select the desired patterns generated by set **B** elements. Any increase in the sets cardinals of **b** or **g** will cause an increase in computing time due to the nonlinear nature and the exponential growth rate of equation (1). This can be noticed by the big O notation [16].

Al-khalil work was as job division or parallelism implementations for the nested loop algorithm, mentioned by the C++ code above, in his time. His work was done by selecting a set of characters at a time, and then permutes each set. To replicate his work, a dynamically populated array list of characters of variable length was used to resemble the set of characters **B**. And a variable option for selecting a set of characters, **G**, from set **B** was used too. The process of selecting **g** characters from **B** set of characters is the same as

implementing [11] this permutation relation $\binom{b}{g} = \frac{b!}{(b-g)!g!}$ in the current time statistical topics. And then Al-khalil permutation was conducted on each selected possible set **g** characters. To demonstrate one instance of the



mentioned process, consider table 1. The first row in table 1 is the array list element. The second row represents the indices of the characters in the array list; the fourth row is the possible selections of **g** characters permutations from **b** characters where each integer is to be replaced with the matching character, as each integer represents an index of a character in the array list. . For example, the set of indices 0123 represents “نجف” And then Al-khalil permutation was conducted on each selected set of characters, shown in the third row of table 1.



ق	س	ن	ج	ف	ب
5	4	3	2	1	0
0123 0124 0125 1234					
1235 2345 0234 0235					
5432 5431 5430 4532					
4321 5321 5320					

Table 1: First row is a set of characters in a dynamically populated array list with cardinal **b** = 6. The second row has integers which represents the indices of the characters in the array list. Third row is dynamically populated array list for ways of permuting **g** = 4 characters out of **b** characters set. Permutations in the third row of table 1 are a job division or parallelizing for the nested loops algorithm to create Al-khalil permutations.

From the result of modeling an equation was derived for the cardinal Al-khalil root words, where geminated characters were counted as one character. The cardinal of Al-Khalil root words, **w**, based on selected set of characters with cardinal, **g**, from a given set of characters with a cardinal, **b**, is given by the following equation.

$$w = f(b, g) = \begin{cases} b! & , \quad g = b \\ \frac{b!}{(b-g)!} & , \quad b > g \end{cases}$$

This equation can be reduced to

$$w = f(b, g) = \frac{b!}{(b-g)!} , \quad b \geq g \quad (2)$$

Equation (2) can be derived mathematically as follows [11]. Let s as the cardinal of the ways for selecting g characters set from b characters set, where $b \geq g$.

$$s = \binom{b}{g} = \frac{b!}{(b-g)!g!}$$

Al-khalil permutation cardinal for the selected g characters is computed as $g!$, factorial of cardinal g . Computing a cardinal resulting from two nested sets cardinals is done by multiplications of both nested sets cardinals, therefore

$$w = sg! = \binom{b}{g}g! = \left(\frac{b!}{(b-g)!g!}\right)g! \\ = \frac{b!}{(b-g)!}$$

The later equation is the same as equation (2), an applicatory examples, consider $b=5$ and $g=5$. The cardinal of Al-khali roots words

$$w = 5!/(5-5)! = 5!/0! = 5 \times 4 \times 3 \times 2 \times 1 = 120 \text{ Words}$$

Now, consider $b=29$ and $g=5$, then the root words cardinal

$$w = 29!/(29-5)! = (29 \times 28 \times 26 \times 27 \times 26 \times 25 \times 24!)/24! = 14,250,600$$

Words

In Equation (3) geminating is considered as one character and the following two cases can be concluded from equation (2)

Case 1: If the characters of the 2,3,4,5 characters root words are selected from separates sets of characters, where the union of any of these sets is empty set, then agglutination "النحت" will not occur among the root words.



Case 2: When 2, 3, 4, and 5 characters root words are selected from the same set of characters, agglutination will occur, that is:

4- characters root word could be an agglutination of two separate sets for 2 characters root words. And 5 characters root word can be an agglutination of 3 characters root word and 2 characters root word.

2.2 Calculation Section

Example 1: Given set of characters $B = \{ف, ش, ب\}$, how many root words, g = 2 characters each can be generated from B ?

Answer: substitute in equation (3) $b = 3$ and $g = 2$

The selected set is:

$$A_s = \{\{ش, ب\}, \{ف, ب\}, \{ف, ش\}\}$$

And the root words are:

$$\{\{ش, ب\}, \{ب, ش\}, \{ب, ف\}, \{ف, ب\}, \{ف, ش\}, \{ش, ف\}\}$$

Example 2: Given a set of characters, $B = \{ب, ق, ش, ف\}$, that have a cardinal $b = 4$, how many 4 characters root words, w , can be generated from this set?

Al-khalil selected set is given by

$$A_s = \{\{ق, ش, ف\}, \{ق, ش, ب\}, \{ب, ف, ق\}, \{ب, ش, ف\}\}$$

The root words generated by the set A_s are shown in the table below

ق ش ف	ق ش ب	ق ف ش	ق ش ف	ق ش ب	ق ش ف
-------	-------	-------	-------	-------	-------



ق ش ب	ب ش ق	ق ب ش	ش ب ق	ب ش ق	ق ش ب
ب ق ب	ق ب ق	ب ق ب	ق ب ق	ب ق ب	ق ب ق
ب ق ب	ق ب ق	ب ق ب	ق ب ق	ب ق ب	ق ب ق
ب ق ب	ق ب ق	ب ق ب	ق ب ق	ب ق ب	ق ب ق

Table 2: Program generated root words for example 2. These are Al-Khalil root words

Example 3: The cardinal of 5 characters root words generated from 5 characters, here b=5 and g=5, substitute

$$w = \frac{5!}{(5-5)!} = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

Example 4: The cardinal of 5 characters root words generated from 29 characters set is given by

$$w = \frac{(29)!}{(29-5)!} = \frac{29 \times 28 \times 27 \times 26 \times 25 \times 24!}{24!} = 14,250,600 \text{ root words}$$

3.1 Root - Derived words, Morphemes and Compression

In programming scripts, word meaning or syntax are not fatally required. Text programming is based on the fundamental concept of a data type named strings. String is a construct of characters sequence that is presented in numerical code for machines processing. Arabic characters are direct implementation of strings processing and the Arabic characters numerical code representation is more than 29 codes, while the Arabic characters are 29 characters. There are some characters with multiple forms, such as the character “أ”, as shown in table 3, and there is a numerical code given for each form. The extra characters that have more than one numerical code must be taken into account when programming Arabic script.

1571	أ
1575	آ
1570	إ
1609	ى

Table 3: shows the numerical code for the form of “أ” character. Visual Studio.net 2008 equivalent integer code is shown in the left column.



An advantage of Al-khalil work on root and derived words morphological relations was taken into account to compress Arabic scripts. The following simplified discussion will administer the basic concept behind a lossless method to reduce the size of files that contains scripts. A set of derived words can be generated from the root words by using what linguists' calls inflecting forms. Arabic language is morphologically rich language. Thus, there are plenty of morphemes. These morphemes exhibit redundancies in the derived words. For example, the semantically derived words "استحسن", "استمسك", "استمتع", "استحب" have in common, the characters "س ت" which exists as a prefix. And, the later words do not have the same length due to the differences in the original root words length. The derived word "صدقهم" has the postfix "هم" or a suffix, which is a morpheme that is added at the end of the root word to form a derivative word. Infix which is mostly long vowels causes variation in the middle of a root word such as "صادق" has one long vowel in the middle. A derivative word may experience a combination of postfix, infix, or prefix morphemes or auxiliary characters. That is all at once, two at time or one at a time. Post fix can be 1, 2, and 3. Additionally, the verb conjugation controls the behavior of the derivative words. Some of the verb forms are "تَفَاعَلَ", "اِسْتَفْعَلَ", "اِنْفَعَلَ". There could be 150 or more derivative words for one root word in Arabic language. Example of derivative words for the word "قصد", are listed below.

"تَقْصِيدٌ", "اَلْاِقْتِصَادُ", ... "يَقْصِدُ", "المَقْصِدُ", "قَصِيدٌ", "مَقْصُودٌ", "قَاصِدٌ", "مَقْصِدٌ", "قَصْدُهُنَّ", "قَصْدُهُمَّ", "قَصِيدَةٌ"

And the root word "صدق" may appear in derivative words as

"أَصْدَقُ", "صَادِقَيْنِ", "صَادِقٌ", "مُصَدِّقٌ", "أَلْصَادِقُ", "يُصَدِّقُ", "صَدِيقٌ", "صَدَقَةٌ", "صَدَقَهُمَا", "صَدَقَهُنَّ", "صَدَقَهُمْ", ...

Due to redundancy of morphemes, morphology is used in script compression especially in dictionaries. It was used in English language to compress text. Other researchers attempted to use it in Arabic as a lossless compression algorithm. [13, 15, 16, 17]. The lossless compression algorithm



is the algorithm that allows full recovery of the compressed data [19]. An example the statement

"استحسن الصادق واستحب اصدقائه الصادقين"

When this statement is encoded by stripping the words from their morphemes, the statement will be compressed to its root words as below.

"حسن صدق و حب 2 صدق"

The number 2 indicate there are two consecutive words of "صدق"; this will reduce the data size in the root words too. Another code, not shown, was a morpheme tagging code so that the original word before encoding can be recovered. Because the character count in root words is less than the derived words, there will be gain in storage size for a given Arabic script. The gain will be for words that has two or more morphemes characters. This is because one code is needed to indentify the morpheme string template as, ("است***") 3 stars represents 3 character root word, which was used to strip the morphemes after correlation of the template with derived words, such as "استحسن". When the Morpheme is removed from a derived word, the word is reverted to its root word. This mechanism can be used in lexical search for word meanings.

Another procedure was tested which was to aid in warning against spelling mistakes in morphemes in an event of typing Arabic text without referencing to dictionaries.

Al-khalil systematic linguistic model makes programming less tedious compared with other words generation approaches, such character elimination "الاختزال" from an existing word or agglutinative "النحت" approach to generate a word from a set of words.

4.1 Geminating, Phoneme and Characters Relations

The Arabic Consonantal System covers a wide range of consonants. This range starts from Laryngeal and Pharyngeal, to labial consonants. Creating words by using sets of abstract alphabets does not require base dialect as in phonemes notations requirements for the International Phonetic Alphabet, IPA. To create a new base word in strictly abstract configurations of alphabets, that is free of short vowels [2], phonetic rules are imposed by pronunciations



phenomenon. For example characters “ع” and “ح” can not be neighbors in a word due to difficulty of pronunciations, which forces the insertion of long vowel or consonant between these two characters that is if these two characters are to be used in a word. Additionally, due to the difficulties of pronunciation, character “هـ” can not precede “ع” as in “هعد”. But, when the order is reversed, character “هـ” can precede “ع” as in the word “عهد”. Ibin Jini, a linguist (ابن جنى) 667-742 AD)[10] stated that there are characters which can not be uttered at once due to nearness of their utterance location in the Arabic consonantal system. The following characters are an example “ع غ ح هـ خ”.

Another apparent grapheme phoneme relations example can be demonstrated by what is called sun and moon characters. The character “ل” in the words “القمر”, “الشمس”, in the first, the character “ل” is not pronounced [6]. The “ش” is voiceless alveolar fricative consonant and it takes shorter period of time to pronounce. The “ل” was graded and forced into geminating with the “ش”, while in the second configurations the “ل” is pronounced. In the first word pattern configuration, the cardinals of the grapheme and the phoneme are not equal. Therefore, the word “الشمس” is not orthogonal word.

No geminating in this word pronunciation “الليل”, but in writing it can be geminated. In some cases geminating is an option. However, geminating can be overridden by pronunciation limitation and meaning variation. For example in the words “حجبت”, “تشدت”, and “مسسنا” characters “د”, “ج”, and “س” can not be geminated in the three stated words. Geminating is related to diphthong which is short vowels and syntax dependent [7, 8, 9]. Ibn Jini, mentioned that when a consonant is uttered twice, the geminating occurs to minimize the work on the tongue [10]. Other word property is when two consonants with zero vowels meet in a derived word [11]. If this occurs in derived word, a conditional elimination or geminating on one of the consonants is stipulated by the Arabic language operators. Geminating does not save a character in the numerical code, as the “م” and “م” both patterns will need two numerical codes. The character with short vowels Fataha over Shaddeh “ي” needs three



numerical codes. Short vowels are the same as long vowels except it takes shorter time to pronounce when used in transition between consonants. The code for Fataha above the characters is different from the numerical code of Fataha below the character.

Characters that can not be pronounced when present next to each other in a word will aid in a dynamically spell checking mechanism without referencing a dictionary, as in the morphemes case, which was discussed earlier in section 3.1.

3.1 Arabic Word Numerical Characters Length.

Comparing Arabic word of zero vowel characters lengths in randomly selected words with other phonetic languages and these are English and French. The derived words in Arabic have on the average the least character count. Even one syllable English words have, on the average more, characters counts than the Arabic words with the same meaning. The comparison was conducted using Google translator for 1000 words, and a sample is given in table 3 below. “Short words are best and the old words when short are best of all.” (Winston Churchill) [18] The fact that the internet is a written language environment short words exhibit less spelling mistake probabilities and require less storage size and transmission time. And, short words are useful for optimizing search engines.

English Word	French Word	Arabic
Mother – 6	Mère -4	أم - 2
Father – 6	Père - 5	أب - 2
Brother -7	Frère -5	أخ - 2
Sister-6	Sœur - 5	أخت-3
Moon -4	Lune -4	قمر -3
Sun -3	Soliel -6	شمس - 3
Room-4	Chamber-7	غرفة - 4



Sky -3	Ciel-3	4- سماء
Strike-6	Grève -5	4- ضربة
Mobile-6	Mobile -6	5- متحرك
Cellular-8	Cellulaire -10	4- خلوي
Mathematics-11	Mathématiques - 13	- رياضيات 7
Physics – 7	Physique - 8	6- فيزياء

Table 3: A comparison of character length for a given words in English, French and Arabic languages, it appears, on the average, the derived Arabic words, including names, have the least character count.

4.1 Conclusions

1- Short words are the most suited words for information storage and transmission because of their numerical equivalent will be small. And, the most used words are short words, which mean there is plenty of redundancy for storage efficiency. Arabic language has been described as rich morphological language, morphemes can be dynamically created and thus information storage and transmission can be efficient in Arabic word.

2- Al-Khalil Linguistic root word Model has a systematic approach which made the Arabic language easier to program specially in lexical. And, it's more compatible with set theory mathematical axioms for information processing.

3- Zero vowel words pattern generation for creating new words in Arabic are better than other approaches, such as phonetic rules implemented by the IPA, due to the absence of base dialects requirements to generate phonemes. Phonemes are social and regional dependent [24]. And, the zero vowels gives the ability to assign the short root words for the most used facts and the longer root words for the least used facts. This generates efficiency in any given language, as efficiency is a measure for getting the most for what is given.



4- Arabic language has a huge reservoir of root words the 4 and 5 root characters, as equation (2) indicates and it was confirmed by programming simulations.

6- Geminating does not save memory. This is because any additional notation requires additional code for logical distinguishing.

7- Long words exhibits agglutination due to the nature of characters repetition in words. Long words may seem to be made of two or more short words that appear as sub words in a single long word.

References

مكتبة مشكاة الإسلامية - كتاب العين للخليل ابن احمد الفراهيدي [1]

الصوتيات العربية الحديثة جامعة سعد دحلب- البليدة - العدد الرابع- 2008- [2]

Forth Edition ISSN: 1112-6426 Pages 22-28, 29-35

[3]http://www.kk.org/thetechnium/archives/2006/02/the_speed_of_in.php

[4] Merriam-Webster <http://www.merriam-webster.com/info/newwords07.htm>

http://www.merriam-webster.com/help/faq/words_in.htm

[6] long-short consonants Geminat:

<http://en.wikipedia.org/wiki/Geminate>

[7]Consonant Mutation: http://en.wikipedia.org/wiki/Consonant_mutation

[8] الإدغام <http://www.reefnet.gov.sy/education/kafaf/Bohoth/Edgham.htm>

[9] <http://www.mazameer.com/vb/t55133.html> الإدغام وأقسامه وأحكامه

[10] تاريخ اللغات عند ابن الجني

<http://www.arabic.almenhaj.net/text.php?linkid=1030>



[11] Roger Nix probability, permutation and combination, chapter 5

<http://www.chem.qmul.ac.uk/software/download/qmc/ch5.pdf>

[11] التقاء الساكنين

<http://www.reefnet.gov.sy/education/kafaf/Bohoth/ElteqaSakenin.htm>

[12] Computer Linguistics :

http://www.coli.unisaarland.de/~hansu/what_is_cl.html

[13] M. Aljlayl, and O. Frieder, “On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming

Approach”, CIKM’02, November 4-9, 2002, M clean, Virginia, USA.

[14] K. Beesley, “Arabic Morphological Analysis on the Internet”.

Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing, Cambridge, 17-18, 1998.

computer Conference and Exhibition, Saudi Arabia, 1986.

[15] S. Khoja. “Stemming Arabic Text”. Lancaster, U.K., Computing Department, Lancaster University,

[16] Big O notation: http://en.wikipedia.org/wiki/Big_O_notation

[17] Combinations and permutation

<http://personal.jccmi.edu/BaarsonMonaG/Math%20145%20Notes%20for%20Class/25-%20%20Permutations%20and%20Combinations.pdf>

[18] Jacob Nielsen: <http://www.useit.com/alertbox/search-keywords.html>

[19] Hoffman encoding: http://en.wikipedia.org/wiki/Huffman_coding

[20] Information theory: http://en.wikipedia.org/wiki/Data_compression

[21] Set theory: <http://www.comp.nus.edu.sg/~fstephan/settheory.pdf>

[22] M. Holz, K. Steffens and E. Weitz (1999).

Introduction to Cardinal Arithmetic. Birkhäuser. ISBN 3764361247.

[23] Algorithmic Information Theory

:<http://docenti.lett.unisi.it/files/4/1/1/43/chaitinalgorithm.pdf>

[24] Sociallinguistics: [http://www.tlumaczenia-](http://www.tlumaczenia-angielski.info/linguistics/sociolinguistics.htm)

[angielski.info/linguistics/sociolinguistics.htm](http://www.tlumaczenia-angielski.info/linguistics/sociolinguistics.htm)

