J. King Saud Univ., Vol. 10, *Comp. & Info. Sci.*, pp. 1-14 (A.H. 1418/1998)

Study of Information-theoretic Properties of Arabic Based on Word Entropy and Zipf's Law

Ibrahim A. Al-Kadi
Electrical Engineering Department, College of Engineering
King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia
Tel. (966-1)467-6796 *Fax*(966-1)467-6757

(Received 11 October 1995, accepted for publication 01 January 1996)

Abstract. Natural languages have very complicated structures but are highly redundant. Statistical studies of a language are extremely important in numerous fields of knowledge, including Education, Linguistics, Computers and Communications.

The aim of this paper is to study the word statistics of the Arabic Language with the intention of estimating the information content of Arabic based on word entropy. Actual statistics of frequencies of Arabic based on a 700,000-word sample will be used to demonstrate the applicability of Zipf's Law to the Arabic Language. Word and letter entropy and redundancy of Arabic are then deduced and compared with corresponding values of English.

1. Introduction

Language is the most basic form of human communication. It is the most natural, and still the most common, way to exchange information and knowledge. Electronic Communications (hereafter referred to simply as communications) is the technology of processing and transferring information. As such, communication is heavily used to transfer language-based information. Natural languages have extremely complicated structures. While different languages may have varying degrees of complexity, they all share highly non-uniform statistical structures. Every language has a high degree of built-in redundancy that enables the speakers of that language to understand it (extract information) even if some parts of the spoken or written message are distorted or changed. One can say that a language has a very robust built-in forward error correcting (FEC) coding, and the human brain has an extremely sophisticated decoder unmatched by any "artificial" FEC scheme ever proposed.

Computational Linguistics is the branch of knowledge concerned with the study of the quantitative properties of language [1-4]. But the study of the statistical properties of a language are also useful in many other fields including education, journalism and mass media, computers and communications. In communications, in particular, the statistical properties of natural languages are extremely important for both source encoding (data compression) and for cryptographic applications [4-9].

A language can be modeled as a discrete Markov source with a finite symbol alphabet and different degrees of inter-symbol dependency. The alphabet set can be individual letters or words in the case of a written language or individual phonemes in a spoken language. This paper is concerned with words in printed Arabic. Most previous studies of information content of Arabic considered only the letters of the Arabic alphabet [1,2,8,10]. The *Information Content* or *Entropy* (***H*** in bit/symbol) of an information source (such as a language) is defined as the statistical average of information per message (sequence of symbols) produced by the source. The entropy of degree ***N*** is given by [5,8,9] :

$$H_N = - \sum_{K=1}^{M} p_k \quad \mathrm{Log}_2 (p_k) \qquad (1)$$

where $p_k = p(X_k^N)$ is the probability of X_k^N, the message consisting of ***N*** consecutive symbols and ***M*** is the number of all possible ($p_k > 0$) messages of length ***N***. Since some messages (sequences of ***N*** consecutive symbols) may never be produced by the source ($p = 0$), it follows that: $M \leq L^N$, where ***L*** is the size of the source's alphabet or the total number of the different symbols available. For example, Arabic has ***L*** = 30 characters (letters or symbols) including space (29 without space); English has ***L*** = 27 (26 without space). In decimal numbering system, there are 10 different symbols (0, 1, 2,...,9), while for binary sources, ***L*** = 2 (***0*** and ***1***).

Entropy is a statistical parameter that measures the degree of uncertainty about the output of the source. Alternatively, entropy is the minimum number of binary digits (bits) per symbol required to code a language in the most optimum code. Equation (1) shows that entropy depends on the statistical properties of the language which is quite complex.

As a first (but highly unrealistic) approximation, if symbols (messages of length one, ***N*** = 1) are assumed to have equal probabilities (uniform distribution), then the zero-degree entropy is:

$$H_o = \mathrm{Log}_2 M > H_N \qquad \text{bit/symbol} \qquad (2)$$

H_0 is an upper limit on entropy. For Arabic with 30 letters (including space), $H_0 = \text{Log}_2 30 = 4.91$ bit/letter (= 4.76 for English). If space is not considered as a letter, then $H_0 = \text{Log}_2 29 = 4.86$ bit/letter (= 4.7 for English) [8].

The first degree entropy, H_1, takes into account the probability of each symbol (letter or word) as a random variable independent of preceding symbols. H_1 can be obtained by setting $N = 1$ in Equation (1). Although still an approximation since languages have complex interdependencies between letters, words and sentences, H_1 is much closer to the real entropy than H_0. The more we take symbol statistical interdependency (the higher value of N or the longer sequences) into account, the better results we obtain. In the limit, the actual language entropy (or information content) is:

$$H = \lim_{N \to \infty} H_N \tag{3}$$

where

$$H \leq H_{N-1} \leq .... \leq H_2 \leq H_1 \leq H_o \tag{4}$$

The equal signs are applicable when symbol inter-dependencies disappear. *Redundancy*, R, of a language (or any information source) is defined as the difference between maximum entropy, H_0 (i.e., when all symbols are independent and of equal probability) and the real entropy H.

$$R = H_o - H \tag{5}$$

Relative Redundancy, r, is ratio of redundancy to the maximum entropy. Hence,

$$r = R / H_o = 1 - H / H_o \tag{6}$$

For the last 40 years, there has been many publications addressing the information-theoretic aspects of different European languages [5,9,11,12]. However, limited results on the information properties of Arabic started to appear only recently [1,2,4,8,10]. In an earlier paper [8], the entropy, H, of Arabic was estimated to be 2.05 bit/letter, which gives redundancy, R, of 2.86 bit/letter and relative redundancy, r, of 58%. The above results are generally higher than the corresponding values of English ($H = 1.23$, $r = 74\%$ [8]). So, Arabic is less redundant (has more information content) than a typical European language such as English. This can be partly explained by the fact that unlike European languages, printed Arabic does not require the use of short vowels as stand-alone letters. Instead Arabic uses special forms of diacritical signs that are marked above or below

consonant letters to indicate short vowels (e.g., ـَ , ـُ , ـِ). In modern standard printed (or hand-written) Arabic texts, these signs are not usually shown but are left to be understood from the context and types of words used. This, of course, could make Arabic a little more prone to errors or misunderstanding by unskilled readers. Diacritical signs may have to be explicitly shown in important documents and in situations where confusion or disputes may arise. The aim of this paper is to study the information content of the Arabic language based on actual word statistics. Next section introduces Zipf's law of word probability distribution in any language. Section 3 presents the main contributions of this paper which include:

- Description of the actual Arabic word statistics employed in the study
- Demonstration of the applicability of Zipf's law applicability to the Arabic language
- Estimation of word and letter entropies for Arabic using two methods--a direct one and one using Zipf's law
- Comparison of the results obtained for Arabic and corresponding results for English

The last two sections present some discussion of the approach and results and final conclusions.

2. Word Entropy and Zipf's Law

An alternative way to finding entropy of a language using statistics of N-gram character strings is to use word statistics [5-8]. To compute word entropy H_w (bits/word), Equation (1) is modified as follows:

$$H_w = - \sum_{K=1}^{M} p(w_k) \, Log_2 [\, p(w_k) \,] \qquad (7)$$

where $p(w_k)$ is the probability of occurrence (or frequency) of word number k. The summation is performed over all word vocabulary of size M. M is the total number of meaningful words (i.e., words where $p(w_k) > 0$). Shannon suggested that letter entropy (bits/letter) be estimated by dividing the word entropy by the average word length [5]:

$$H = H_w / N_w \qquad (8)$$

where N_w is the average number of letters per word. N_w for Arabic was estimated to be 5.17 letters/word (with space) or 4.17 (without space) [6,8]. The corresponding values

for English are 5.5 and 4.5 letters/word respectively [5,8]. The fact that Arabic words are generally shorter than English words can again be explained by absence of short vowels in printed Arabic as discussed earlier.

In order to compute word entropy for English, Shannon made use of Zipf's Law to estimate the probability of words in English [5]. The American linguist G. K. Zipf [5,13,14] proposed a simple law to estimate the frequency of occurrence of various words in any language. Zipf's Law is given by:

$$p(w_k) \approx A / k \qquad (9)$$

where A is a constant depending on the language under consideration, and k is the order of the word in terms of its frequency. So, if the words in a language are ordered in decreasing order of frequency of occurrence (from the most frequent to the least frequent), then, on the average, the second most frequent word (the third, the fourth...or the k^{th}) occurs half (one-third, one-fourth, ...or one-k^{th}) as many times as the most frequently used word. It was found that Zipf's approximation holds remarkably well for many quite different languages including English, Yiddish, Old German and Norwegian [5,9,11]. Figure 1 shows the fit of Zipf's Law to actual word frequencies for English [5,11].

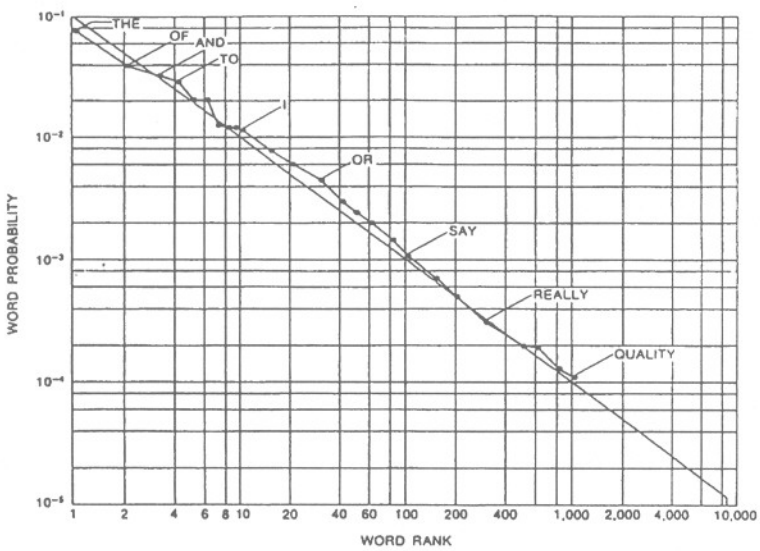

Fig. 1. Zipf's law and word probability distribution for English.

Based on actual Arabic word statistics, this paper will show that Zipf's Law is also valid for Arabic (See Next Section). To the best of the author's knowledge, no other work has been done before to test the applicability of Zipf's Law to Arabic.

The language-dependent constant A is usually estimated from curve fitting of actual word statistics of the language under consideration and a straight line on a log-log scale (For English, A was found to be about 0.1). Another systematic and more accurate method to estimate the constant A will be presented in this paper (see next section.). Since $p(w_k)$ is a probability function then:

$$1 = \sum_{K=1}^{M} p(w_k) \approx \sum_{K=1}^{M} A / k \qquad (10)$$

So based on A, there is a specific M (total number of words under consideration) for the sum to equal exactly one. For English with $A = 0.1$, M was found to be 12366 [12]. In his famous paper [5], Shannon had made a computational slip that gave erroneous results (M = 8727 and word entropy = 11.82 bit/word) [5]. Such an error remained undetected for more than 20 years [12]. To find word entropy, Equations (7), (9) and (10) are used to get:

$$H_w = - \sum_{K=1}^{M} A / k \, [\, \mathrm{Log}_2 (A / k) \,] \qquad (11)$$

$$H_w = A \sum_{K=1}^{M} [\, \mathrm{Log}_2 (k) / k \,] - \mathrm{Log}_2 (A) \qquad (12)$$

The English word entropy (with $A = 0.1$ and $M = 12366$) is found from Equation (12) to be 9.71 bit/word. Using Equation (8), the letter entropy of English is 2.16 bit/letter based on 4.5 letter per word without space [9,11,12] (or 9.71/5.5 = 1.77 bit/letter with space).

3. Arabic Word Entropy

Studies on the Arabic word statistics were performed by different researchers [1-3,7,15]. As far as the author knows, however, the work of Abdo [3] is the best and most reliable linguistic study. Abdo's sources included four earlier lists containing about 715,000 words of the most frequent Arabic words taken from different modern printed Arabic sources such as newspapers, elementary reading textbooks and general books in

diverse fields including poetry, literature, history, economics, education and sociology. Abdo's list contains 3025 words that are the most frequent in printed Arabic together with their number of occurrences in the four source lists. Table 1 is a list of the 50 most frequent words in Arabic which account for about 30% of all printed Arabic words. The first 100 words accounts for more than 36% of total words in Arabic text. The 3025 words of Abdo's list account for nearly 83.4% of all printed Arabic words. All remaining words, not listed in Abdo's list, account for only about 16% of words in normal Arabic texts.

Abdo's statistics have been used by the author and his student to study the information-theoretic properties of the Arabic language [6,7]. In particular, the probabilities of the 3025 most frequent Arabic words were used to test the applicability of Zipf's Law to Arabic. The author is not aware of any similar study for Arabic.

Figure 2 is a plot (on a log-log scale) of Arabic word actual probabilities, $p(w_k)$, versus word rank, k. The best straight line that fits the actual probability distribution is a line (A/k) with A =0.11.

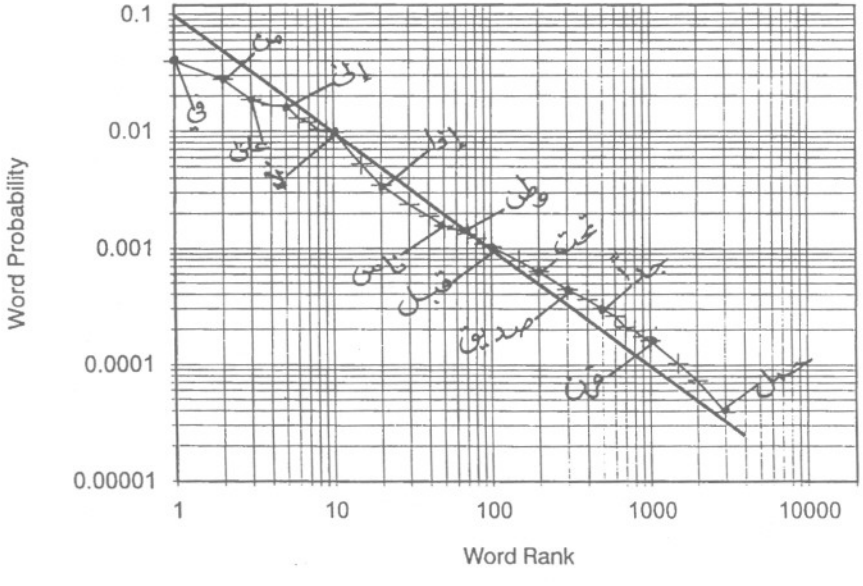

Fig. 2. Application of zipf's law to the Arabic language (word probability versus word rank).

A more accurate method to find the constant A is proposed here as follows: Since the 3025 most frequent Arabic words constitute 83.4% of all Arabic words, then:

$$\sum_{k=1}^{3025} p(w_k) = \sum_{k=1}^{3025} A / K = 0.833509 \quad (13)$$

Using MATHCAD® mathematical package on a PC, the A was found to be $A = 0.09700936$.

Table 1. A list of the 50 most frequent Arabic words and their probabilities of occurrence

Rank	Word	Frequency %		Rank	Word	Frequency %
١	في	٣,٩٦٥٥		٢٧	أو	٠,٢٩١٠
٢	من	٢,٧٦٦٣		٢٨	حتى	٠,٢٦٣١
٣	على	١,٨٥٧٢		٢٩	لكن/ لكن	٠,٢٤٧٨
٤	إن – أنَ	١,٦٥٨٢		٣٠	كثير	٠,٢٣٨٢
٥	إلى	١,٦٥١٥		٣١	من	٠,٢٣٤١
٦	كانَ	١,٢٩٦٦		٣٢	غير	٠,٢٢٨١
٧	هذا / هذه	١,١٨٧٧		٣٣	نفس	٠,٢٢٢٥
٨	أن	١,٠٢٨٧		٣٤	ماذا/ لماذا	٠,٢١٨٢
٩	الذي/ التي	٠,٩٨٩٠		٣٥	الله	٠,٢٠٩٤
١٠	لا	٠,٩٧٦٨		٣٦	يا	٠,٢٠٥٦
١١	ما	٠,٩٠٧١		٣٧	حكومة	٠,١٩٤١
١٢	قال	٠,٨٩٠٠		٣٨	إلا	٠,١٩٤٠
١٣	قد	٠,٦٦٣٢		٣٩	أب	٠,١٩٠٣
١٤	عن	٠,٦٠٤٣		٤٠	ملك	٠,١٨٨٥
١٥	ذلك / تلك	٠,٥٢٢٧		٤١	أخذ	٠,١٨٨٤
١٦	كُل	٠,٤٩٩٧		٤٢	بعض	٠,١٨٧٣
١٧	لم	٠,٤١٤٣		٤٣	كبير	٠,١٨٠٤
١٨	ثم	٠,٣٦٧٤		٤٤	أول	٠,١٧٩٩
١٩	هو	٠,٣٥٠٨		٤٥	شيء	٠,١٧٦٦
٢٠	إذا	٠,٣٤٤٩		٤٦	عند	٠,١٦٢٩
٢١	بين	٠,٣٤٢٩		٤٧	بلاد	٠,١٦٢٢
٢٢	رأى	٠,٣٣٧١		٤٨	جميع	٠,١٥٨٩
٢٣	يوم	٠,٣٣٢٢		٤٩	سيد	٠,١٥٧٢
٢٤	مع/معاً	٠,٣٠٧٣		٥٠	ناس	٠,١٥٦١
٢٥	هي	٠,٣٠٦٠		First 50 Words		٪٢٩,٧١
٢٦	بعد	٠,٣٠٤١		First 100 Words		٪٣٦,٠٤

Next, we turn our attention to finding an estimate of the total number of Arabic words that are used in any text (100% of all useable words). Since Σ_1^{∞} 1/n is not finite, we need to find the maximum Arabic vocabulary size M such that the sum $A(1 + 1/2 + 1/3 + 1/4 + ...+ 1/M)$ is as close to unity as possible. By iteration using MATHCAD®, ***M*** was found to be ***16832*** words which gives:

$$\sum_{k=1}^{16832} 0.09700936 / k = 1 - 10^{-7}$$

This means that with probability of about one (near certainty), any word we pick at random from any Arabic text will be one of the most frequent ***16832*** Arabic words. To recognize all words in any Arabic text, one has to know ***16832*** words (12366 for English). Finally, with A = 0.09700936 and ***M = 15543***, we use Equation (12) to find word entropy of Arabic as:

$$H_w = 9.98196391735 \quad \text{(bits/words)} \tag{14}$$

Letter entropy of Arabic is ***1.9307*** bit/letter including space (9.98196 / 5.17) or ***2.3938*** bit/letter (9.98196 / 4.17). Based on the above results, redundancy, ***R***, of Arabic is ***2.9762*** bit/letter (***2.4642***) and relative redundancy, ***r***, is ***61%*** (***51%***) with (without) space.

Another more straightforward method to estimate Arabic word entropy is to directly use Equation (7), with actual word probabilities instead of Zipf's law approximated probabilities. A computer program was written and used to convert frequencies of occurrence of the 3025 most common words as given by Abdo into word probabilities, $p(w_k)$ ($k = 1, 2, 3.....$, 3025), and then compute word entropy, H_w, letter entropy, ***H***, redundancy, ***R***, and relative redundancy, ***r***, using Equations (7), (8), (5) and (6) respectively. The results obtained are :

$$H_w = 9.6998 \quad \text{bit/word} \tag{15}$$

$$H = 1.88 \quad \text{bit/letter} \quad \text{(2.33 without space)} \tag{16}$$

$$R = 3.03 \quad \text{bit/letter} \quad \text{(2.53 without space)} \tag{17}$$

$$r = 62\% \quad \text{(52\% without space)} \tag{18}$$

These results agree quite well (within 2%) with the results obtained above using Zipf's law in Equations (12). To the best of the author's knowledge, there are no published work describing the direct method presented above, and no similar results for English or any other language using the direct method are available.

Recall that for English, Zipf's law results are 9.71 bit/word, 1.77 bit/letter with space and 2.16 bit/letter without space. Accordingly, redundancy of English is 2.99 bit/letter with space (2.54 without space) and relative redundancy is 62.8% (54% without space).

4. Conclusion

In this paper, actual Arabic word statistics were used to test Zipf's Law applicability to Arabic. The results have demonstrated that Zipf's Law holds quite well for Arabic. This adds more credibility to the correctness of Zipf's Law which was earlier found to hold for many different other languages. In addition, we have proposed a simple systematic method to estimate the value of the language constant A in Zipf's Law. The proposed method was used to find the constant A for the Arabic language. Zipf's Law was used to estimate the vocabulary size of all words in printed Arabic today. A total of around 16850 Arabic words constitute all words in common use.

Relevant formulas (based on Zipf's Law) have also been developed and used to compute Arabic word and letter entropies. Arabic word entropy is about 9.982 bit/word. Arabic letter entropy calculated from word entropy is 1.93 bit/letter including spaces between words, or 2.39 bit/letter not counting space as a letter. These values are in close agreement with Arabic letter entropy computed from individual letter statistics (2 and 2.4 bit/letter, respectively [8].)

An alternative direct method based on the use of actual word probabilities was also used to re-compute the same information-theoretic parameters . Results obtained using both approaches, the direct method and Zipf's law, are in close agreement. To summarize the different results derived in this paper, Table 2 is a tabulation of the important information properties of Arabic as in the paper together with corresponding value for English for comparison.

Finally, two notes of caution, concerning the estimation of word entropy, are in order:

i) Equation (7) implies that words are independent. Words in typical Arabic (or any other language) texts are not independent. The equation gives the first order word entropy. A more accurate, but for more complex, method is to use Equation (1),

Table 2. Summary of the information-theoretic properties of Arabic and English

Parameter	Arabic		English	
	With space	No space	With space	No space
I - General				
1.1 - Number of letters	30	29	27	26
1.2 - Average word length N_w (ltr/word)	5.17	4.17	5.5	4.5
1.3 - Maximum Entropy, H_0 (bit/letter)	4.9069	4.85799	4.7542	4.7004
2 - Letter statistics [8]				
2.1 - Entropy, H (bit/letter)	2.05	2.49	1.23	1.5
2.2 - Redundancy, R (bit/letter)	2.8569	2.36799	3.5242	3.2004
2.3 - Relative Redundancy, r	58.2 %	48.6 %	74.1 %	68.1 %
3 - Actual word statistics				
3.1 - Number of Individual Words	3025 words		RESULTS BASED ON	
3.2 - Size of Statistics Sample	715,000 words		ACTUAL WORD	
3.3 - Word Entropy, H_w (bit/word)	9.699786		STATISTICS ARE NOT	
3.4 - Entropy, H (bit/letter)	1.8762	2.3261	AVAILABLE FOR	
3.5 - Redundancy, R (bit/letter)	3.0307	2.5319	ENGLISH	
3.6 - Relative Redundancy, r	61.8 %	52.1 %		
4 - Zipf's law results				
4.1 - Language Constant, A	0.09700936		0.1	
4.2 - Number of words in language	16832		12366	
4.3 - Word Entropy, H_w (bit/word)	9.9819631735		9.71	
4.4 - Entropy, H (bit/letter)	1.9307	2.3938	1.77	2.16
4.5 - Redundancy, R (bit/letter)	2.9762	2.4642	2.9842	2.5403
4.6 - Relative Redundancy, r	60.6 %	50.7 %	62.8 %	54.0 %

with X_K^N representing N consecutive words. This would require the use of n^{th} order word statistics which are not generally available. Equation (7) is an over-estimation of word entropy.

ii) Equation (6) implies that letters within a word are independent (i.e., the probability of a word is the product of the individual probabilities of each of the N_w letters within the word.) Letter entropy computed from Equation (8) is an underestimate.

The overall result, however, is closer to reality since the errors in (i) and (ii) are in opposite direction. Equation (8) underestimates letter entropy (Note ii) based on the use of an over-estimated word entropy (Note I).

References

[1] Mackay, P. (Ed.). *Computers and the Arabic Language*. N.Y. : Hemisphere Publishing Co., 1990.

[2] Descout, R. (Ed.). *Applied Arabic Linguistics and Signal & Information Processing*. N.Y.: Hemisphere Publishing Co., 1987.

[3] Abdo, D. A. *Frequent Words in the Arabic Language*, (in Arabic). Riyadh University Press, Riyadh, Saudi Arabia, 1979.

عبده، داود عطية. *المفردات الشائعة في اللغة العربية*. الرياض: مطبوعات جامعة الرياض ، ١٣٩٩ هـ / ١٩٧٩ م .

[4] King Abdulaziz Public Library (Ed.). *Proceedings of Symposium on Using Arabic Language in Information Technology* (8-12 Dhu Al Qadah 1412 A.H. / 10-14 May 1992), (in Arabic), King Abdulaziz Public Library Press - Authentic Works Series (4), Riyadh, 1993.

مكتبة الملك عبدالعزيز العامة. *السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات* (٨-١٢ ذي القعدة ١٤١٢ هـ /١٠-١٤ مايو ١٩٩٢ م)، *سلسلة الأعمال المحكمة* (٤). الرياض: مطبوعات مكتبة الملك عبدالعزيز العامة، ١٤١٤هـ/١٩٩٣م.

[5] Shannon, C. E. "Prediction and Entropy of Printed English." *Bell System Technical Journal*, 30 (1951), 50 - 64.

[6] Al-Mas'oud, F. A. "*Cryptography and Information : Case Study on the Arabic Language*, (in Arabic)." Graduation Project (Supervised by Dr. I. Al-Kadi), Electrical Engineering Dept., College of Engineering, King Saud University, 1990.

المسعود، فهد أحمد. *التشفير و المعلومات – دراسة حالة اللغة العربية*، مشروع تخرج (إشراف د. إبراهيم القاضي). الرياض: قسم الهندسة الكهربائية، كلية الهندسة، جامعة الملك سعود، جمادى الثاني ١٤١٠ هـ / يناير ١٩٩٠ م .

[7] Al-Obaid, A. I. "*Data Compression in the Case of the Arabic Language with Some Applications in Cryptology*, (in Arabic)." Graduation Project (Supervised by Dr. I. Al-Kadi and Dr. A. Al-Jabri), Electrical Engineering Dept., College of Engineering, King Saud University, January 1992.

العبيد، عبد الله إبراهيم. *ضغط المعلومات في دراسة اللغة العربية مع بعض الاستخدامات في علم التعمية* ، مشروع تخرج (إشراف د. إبراهيم القاضي و د. عبد الرحمن الجبري). الرياض: قسم الهندسة الكهربائية، كلية الهندسة، جامعة الملك سعود، رجب ١٤١٢هـ / يناير ١٩٩٢م .

[8] Al-Kadi, I. A. "Cryptology and Data Security: Cryptographic Properties of Arabic." (in Arabic). *Arab Gulf Journal of Scientific Research*, 11, No. 3 (1993), 457-485.

القاضي، إبراهيم. " الخصائص المعلوماتية والتعموية للغة العربية " ، *مجلة الخليج العربي للبحوث العلمية*،١١ ، العدد ٣ ، رجب ١٤١٤هـ (ديسمبر ١٩٩٣م) ، ٤٥٧ – ٤٨٥ .

[9] Welsh, D. *Codes and Cryptography*. Oxford Science Publications, Clarendon Press, Oxford, UK, 1988.

[10] Wanas, M. A. *et al*. "First, Second and Third Order Entropies of Arabic Texts." *IEEE Transactions on Information Theory*, Vol. IT-22, 1976, p. 123.

[11] Miller, G. A. *Language and Speech*. San Francisco : Freeman and Co., USA, 1981.

[12] Yavus, D. "Zipf's Law and Entropy." *IEEE Transactions on Information Theory*, 20 (1974), p. 650.

[13] Zipf, G. K. *The Psycho-Biology of Language*. Houghton Mifflin, 1935.

[14] Zipf, G. K. *Human Behavior and the Principle of Least Effort*. Addison Wesley Press, 1949.

[15] Al-Fairozabadi, M. M. Y. (d.1414 A.D.), *"Basa'er Theoey al-Tamyyez Fi Lata'if al-Ketab al-Aziz*. (in Arabic)." Edited by M. A. Al-Najjar, Part 1, Scientific Library, Beirut.

الفيروز آبادي، مجد الدين محمد بن يعقوب. (المتوفى سنة ٨١٧ هـ). *بصائر ذوي التمييز في لطائف الكتاب العزيز* ، تحقيق محمد علي النجار، الجزء الأول . بيروت : المكتبة العلمية، ١٤١٤هـ.

[16] Mousa, A. H. "Use of Computers in the Study of Words in the Holy Qur'an." (in Arabic) Kuwait: A*alam al-Fikr*, 12, No. 4, (1982), 153-194.

موسى، علي حلمي. "استخدام الحاسب الإلكتروني في دراسة ألفاظ القرآن الكريم". الكويت،وزارة الإعلام، *مجلة عالم الفكر*، ١٢، العدد٤، يناير/فبراير/مارس، (١٩٨٢)، ١٥٣ — ١٩٤.

دراسة الخواص المعلوماتية للغة العربية: إنتروبيا الكلمات وقانون زيبف

إبراهيم عبد الرحمن القاضي

قسم الهندسة الكهربائية، كلية الهندسة ، جامعة الملك سعود، ص.ب ٨٠٠،
الرياض ١١٤٢١، المملكة العربية السعودية

(قدّم للنشر في ١٩٩٥/١٠/١١م؛ وقبل للنشر في ١٩٩٦/١/١م)

ملخص البحث. تمتاز اللغة العربية (وباقي اللغات الطبيعية) بأنها ذات تركيب إحصائي بالغ التعقيد، ولكنها أيضاً ذات تكرارية عالية تتيح للناطقين بها معرفة المعاني المقصودة حتى لو فقدت أو تغيرت بعض أجزاء الكلام . وتشكّل دراسة الخواص الإحصائية للغة أهمية قصوى في كثير من العلوم الإنسانية والهندسية مثل التربية، تعليم اللغات، اللسانيات، المعلومات، الحواسيب، الاتّصالات، و معالجة الإشارات .

يرمي هذا البحث إلى دراسة الخواص الإحصائية للكلمات العربية المطبوعة بهدف تقدير معدّل المعلومات في اللغة العربية . وسيتم استخدام إحصائيات للكلمات العربية الأكثر شيوعاً والمأخوذة من عيّنات نصوص طويلة و متنوعة في مختلف مجالات المعرفة تضم أكثر من ٧٠٠,٠٠٠ كلمة . وستستخدم هذه الإحصائيات في إثبات صحة قانون "زيبف" حول التكرار النسبي للكلمات في اللغة الإنسانية الطبيعية ، والذي ينص على أن احتمال ورود أي كلمة في لغة ما يتناسب مع ترتيب هذه الكلمة في اللغة. وسيتم في النهاية تقدير معدّل المعلومات (الإنتروبيا) للكلمات والحروف العربية بطريقتين : إحداهما مباشرة والأخرى باستخدام قانون "زيبف" . كما ستتم مقارنة النتائج للغة العربية مع القيم المناظرة للغة الإنجليزية.