World Applied Sciences Journal 21 (Special Issue of Studies in Language Teaching and Learning): 168-173, 2013
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2013.21.sltl.2151

A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty

[1]Nuraihan Mat Daud, [2]Haslina Hassan and [3]Normaziah Abdul Aziz

[1]Kulliyyah of Languages and Management
[2]Kulliyyah of Islamic Revealed Knowledge and Human Sciences
[3]Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia

Abstract: The present study is aimed at designing a formula for estimating the difficulty of reading Arabic texts. Flesch, Gunning Fox and Dale-Chall are some of the formulae that have been used for measuring English texts difficulty. Some of them have been automated making it easy for users to check the readability level of a particular text. A few scholars have attempted to come up with a readability formula for Arabic, but none has been automated. This study is thus conducted to find the formula that would make it possible for users to measure the difficulty level of Arabic texts online. This will greatly help in materials selection for reading comprehension and testing. This paper will present the prototype of a readability formula which is based on a corpus for estimating the difficulty of Arabic written documents.

Key words: Readability · Corpus · Arabic · Reading · Writing · Text

INTRODUCTION

When selecting materials for a textbook or for an examination, one of the issues that would have to be addressed is the suitability of the reading levels of texts. Research has shown that personal judgments about text difficulty are not valid indicators of reading level [1, 2, 3] and comprehension can be difficult if the difficulty level of the texts is higher than the learners' reading level [4]. Several readability formulae have been proposed to estimate a text reading ease [5]. "A readability formula is a mathematical equation that is applied to prose texts to predict how difficult the text will be for a given group of readers" [6]. It measures the appropriateness of texts to a particular group of readers. Among the popular readability formulae are the Flesch formula, Dale-Chall, Gunning Fog Index, Fry Readability Graph, McLaughlin's SMOG and the FORCAST formulae. Readability is widely used in education to develop materials for language teaching, to select suitable textbooks for students, to help teachers' select suitable reading materials for their students and to assess the difficulty level of texts used in language testing.

A number of studies have been done with regard to text readability. In 1953, Wilson Taylor created a cloze test to estimate the readability level of a text by measuring an individual's understanding of a given text. In this test, the intended audience is given a text with missing words at regular intervals (usually every fifth word) and then he/she is asked to fill in the blanks. The percentage of correct words is calculated to produce the cloze score. If a reader fills in the missing words correctly, this indicates that he/she understands the text. The cloze scores can categorize the reader into three reading levels: independent, instructional and frustrational reading levels.

[7] applied the readability formula in their study on cloze procedure as a test of plagiarism. They found that documents that are difficult to read (plagiarized or paraphrased) yielded significantly lower cloze scores than easier to read documents.

[8], [9] and [10] used the Flesch Reading Ease Index in their study to analyze the predicted readability of intermediate accounting texts. All found little or no significant differences among the intermediate accounting

Corresponding Author: Nuraihan Mat Daud, Kulliyyah of Languages and Management, International Islamic University Malaysia, Kuala Lumpur, Malaysia.

texts that they analyzed. The study finds no compelling evidence, in terms of readability, to choose any one of the texts over another.

The mean of articles from the AAOS website was studied by [11] using Flesch-Kincaid to find the readability of online patient education materials. Only 10 (2%) of the articles had the recommended readability level of sixth grade or lower. The articles readability did not change with time. The findings suggest that the majority of the patient education materials available on the AAOS Web site have readability scores that may be too difficult for comprehension by a substantial portion of the patient population. [12] made a similar finding in their study. They found that the readability level of the online mental health brochures that they investigated was higher than the 8th grade level recommended for educational material by the U. S. Department of Education.

Although the formula is widely used on texts written in English, little attention has been paid to its use in Semitic languages such as Arabic. There is thus a need to generate a readability formula for Arabic to assist teachers, test-setters and textbook writers in choosing the appropriate texts to serve their purpose.

Available Readability Formula for Arabic: Two formulae have been produced to measure Arabic text readability, namely Dawood and Al Heeti formulae [13]. Dawood formula includes five readability features, which include: average word length, average sentence length, word frequency, percentage of nominal clauses and percentage of definite nouns, wheras Al Heeti formula includes one factor only i.e AWL (Average Word Length) = (AWL*4.414)-13.468

Both of the available formulae look at either the number of words, syllables and sentences when developing a formula for assessing text difficulty. This, however, did not take into account the fact that some words are used more often than others. Words that are frequently used are usually easier that those that are hardly used. The high frequency words are often easier than the low frequency words. Hence, it is also important to differentiate the frequency of usage in determining a text level of difficulty. Another important issue to be addressed is the ease of use. The available formulae would have to be calculated manually which can be time-consuming and laborious. This study attempts to automate the process to make it more user-friendly to the personnel concerned.

Objectives and Method of Study: This study aims to generate another formula for measuring Arabic texts level of readability. It is proposing the use of a corpus as it is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language [14]. The term 'corpus' is derived from the Latin word for 'body'; hence any body of a text is a corpus. The contemporary corpus is digitized and stored electronically for easy access. Its availability allows for linguistics analysis using text analysis software [15].

There exists a long list of existing Arabic corpora as listed by [16] in Table 1:

However, not all of the above corpora are easily accessible and freely available. Among the Arabic corpus available on the Internet are KACSTAC and IIUMAC. The former is a general corpus where the sources are derived from magazines, books, newspapers, referred journals, dissertations, government circulation, school curriculums, newswire and the Internet. While the latter is a specialized academic corpus, which is an Arabic corpus, based in the International Islamic University Malaysia. It is accessible online through its website: http://efolio.iium.edu.my/arabicconcordancer.

For the purpose of this study, King Abdulaziz City for Science and Technology Arabic Corpus (KACSTAC) was utilized since it reflects a more general use of the Arabic language. KACSTAC consists of 739,119,011 words with 746, 4396 type token ratio (non-repeated words). Figure 1 displays a screenshot of the KACSTAC corpus that was used in the study.

In the KACSTAC corpus, the word with the highest number of frequency is ranked last. Hence, the easiest word will have the highest number. In this study, the ranking in the corpus is reversed so that the easiest number is ranked the first and so on. The difficulty level based on this new ranking is taken into consideration when the mean is computed.

Formula Development: Most of the available formulae look at either the number of words, syllables and sentences when developing a formula for assessing text difficulty. However, for Arabic it can be argued that a higher number of words in a text does not mean that the text is more difficult. Texts with simple sentences and higher frequency words would be easier to read than texts with complex sentences and low frequency words. Similarly a shorter syllable does not mean that it is easier since many Arabic words consist of three syllables, example, kataba. These were taken into account when drawing the formula for Arabic texts.

When the KACSTAC corpus is used, the average frequency is calculated out of the total number of words in a sentence. For example: أحب رسول ال

Table 1: Existing Arabic Corpora [16]

Name of Corpus	Source	Medium	Size	Purpose	Material
Buckwalter Arabic Corpus 1986-2003	Tim Buckwalter	Written	2.5 to 3 billion words	Lexicography	Public resources on the Web
Leuven Corpus (1990-2004)	Catholic University Leuven, Belgium	Written and spoken	3M words (spoken: 700,000)	Arabic-Dutch/ Dutch-Arabic learner's dictionary	Internet sources, radio and TV, primary school books
Arabic Newswire Corpus (1994)	University of Pennsylvania LDC	Written	80M words	Education and the development of technology	Agence France Presse, Xinhua News Agency and Umma Press
CALLFRIEND Corpus (1995)	University of Pennsylvania LDC	Conversational	60 telephone conversations	Development of language identification technology	Egyptian native speakers
NijmegenCorpus (1996)	Nijmegen University	Written	Over 2M words	Arabic-Dutch/ Dutch-Arabic dictionary	Magazines and fiction
CALLHOME Corpus (1997)	University of Pennsylvania LDC	Conversational	120 telephone conversations	Speech recognition produced from telephone lines	Egyptian native speakers
CLARA (1997)	Charles University, Prague	Written	50M words	Lexicographic purposes	Periodicals, books, internet sources from 1975-present
Egypt (1999)	John Hopkins University	Written	Unknown	MT	A parallel corpus of the Qur'an in English and Arabic
Broadcast News Speech (2000)	University of Pennsylvania LDC	Spoken	More than 110 broadcasts	Speech recognition	News broadcast from the radio of voice of America.
DINAR Corpus (2000)	Nijmegen Univ., SOTETEL-IT, co-ordination of Lyon2 Univ	Written	10M words	Lexicography, general research, NLP	Unknown
An-Nahar Corpus (2001)	ELRA	Written	140M words	General research	An-Nahar newspaper (Lebanon)
Al-Hayat Corpus (2002)	ELRA	Written	18.6M words	Language Engineering and Information Retrieval	Al-Hayat newspaper (Lebanon)
Arabic Gigaword (2002)	University of Pennsylvania LDC	Written	Around 400M	Natual language processing, information retrieval, language modelling	Agence France Presse, Al-Hayat news agency, An-Nahar news agency, Xinhua news agency
E-A Parallel Corpus (2003)	University of Kuwait	Written	3M words	Teaching translation and lexicography	Publications from Kuwait National Council
General Scientific Arabic Corpus (2004)	UMIST, UK	Written	1.6M words	Investigating Arabic compounds	http://www.kisr.edu.kw/science/
Classical Arabic Corpus (CAC) (2004)	UMIST, UK	Written	5M words	Lexical analysis research	www.muhaddith.org and www.alwaraq.com
Multilingual Corpus 2004	UMIST, UK	Written	11.5M words (Arabic 2.5M)	Translation	IT-specialized websites-computer system and online software help-one book
SOTETEL Corpus	SOTETEL-IT, Tunisia	Written	8M words	Lexicography	Literature, academic and journalistic material
Corpus of Contemporary Arabic (CCA) 2004	University of Leeds	Written and spoken	Around 1M words	TAFL	Websites and online magazines
DARPA Babylon Levantine Arabic Speech and Transcripts (2005)	University of Pennsylvania LDC	Spoken	About 2000 telephone calls	Machine translatioon, speech recognition and spoken dialogue system	Fisher style telephone speech collection

Table 2: Ranking of Words in a Sentence

Words	Word ranking as in KACSTAC
كان	21
أسد	5430
قوى	3022
الجسم	2375
طيب	4350
القلب	1103
يعيش	2166
وسط	704
مملكة	6592*
سعيدا	242*

Table 3: Average of Word Frequency Count for Each Sentence

Level	Total reversed ranking of each word in a sentence/ no of words per sentence
Advanced يستطيع المشي على الرمل	$\frac{1152+9640+3+9049}{4}=\frac{19844}{4}=4961$
Intermediate كان أسد قوي الجسم	$\frac{21+5430+3022+2375}{4}=\frac{10848}{4}=2712$
Beginners ونصلي على سيدنا ومولانا محمد	$\frac{23+2591+3+7339+23+3227+34}{7}=\frac{13240}{7}=1891$

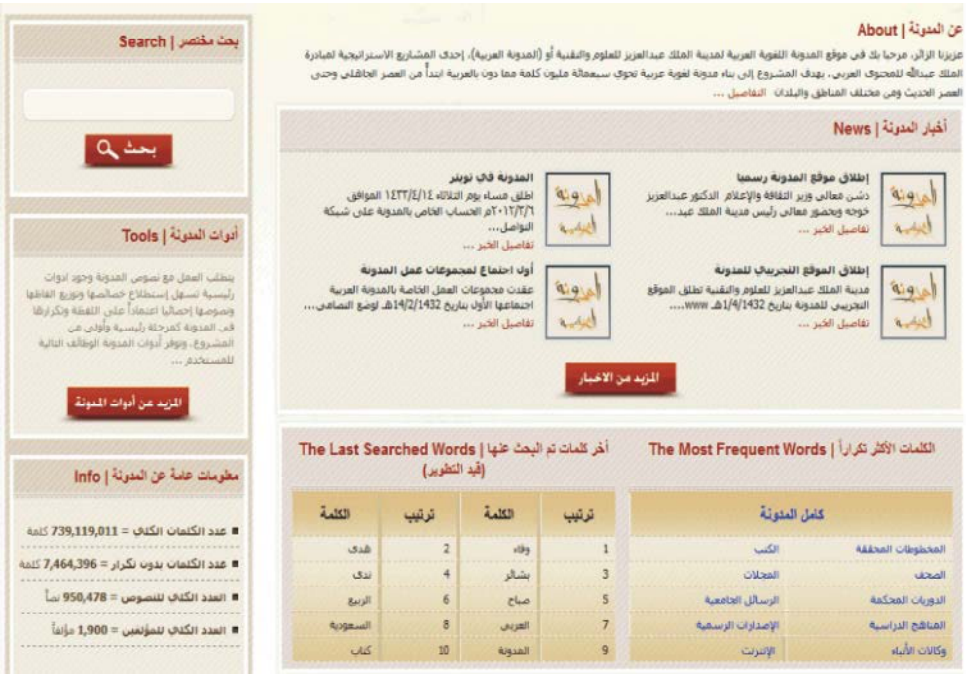

Fig. 1: Screen shot of KACSTAC Corpus

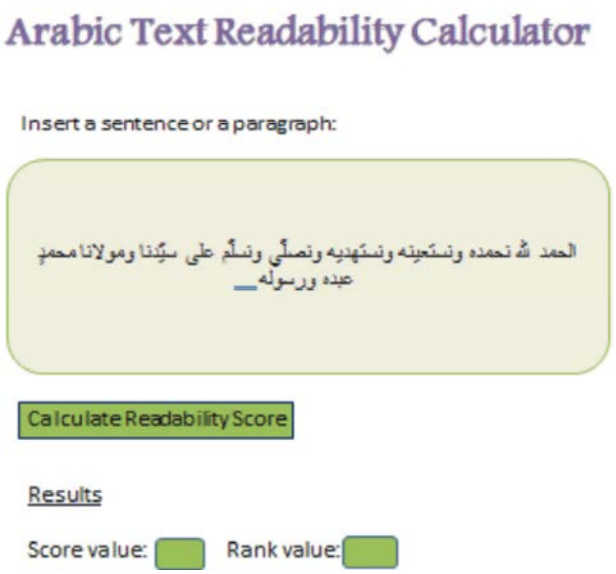

Fig. 2: A Screen-shot of Arabic Text Readability Prototype System

The word أحب ranked 2361 in the KACSTAC corpus رسول is ranked 82 and the word الله is ranked 6 in the corpus. When calculated, the average of this would be as follows:

The total reversed ranking of each word in a sentence/number of words per sentence is:

(6 + 82 + 2361) / 3 = 816

The same principle is applied on an intermediate level text as in the sentence below (ranking in Table 2).

كان أسد قوى الجسم , طيب القلب , يعيش وسط مملكته سعيدا.

$$\frac{\text{Total reversed ranking}}{\text{No of words per sentence}} = \frac{21 + 5430 + 3022 + 2375 + 4350 + 1103 + 2166 + 704 + 6592 + 242}{10} = 2{,}600.5$$

The following sentences illustrate further how the calculation is done using word frequency count:

The next step in this study is to automate the calculation for text difficulty based on the KACSTAC corpus to make it available online. This is done using the above formula. To date, the system is still in its prototype version. The following diagram is a screen-shot of the automated system.

CONCLUSION

The proposed formula for calculating text difficulty can be easily understood by a language teacher as the argument is based on a language formula. With this knowledge, teachers can select teaching materials according to their students' level of proficiency. The same formula can be applied when they want to decide what texts to be included in an examination question. This formula, however, can only be used to compare the estimated level of difficulty of a text to another. Further research needs to be conducted to set the range for each learning level (for beginners, intermediate, advanced).

REFERENCES

1. Burke, Victoria and Greenberg Daphne, 2010. Determining Readability: How to Select and Apply Easy-to-Use Readability Formulas to Assess the Difficulty of Adult Literacy Materials, Adult Basic Education and Literacy Journal, 4(1): 34-43.
2. Hamilton, C. and M. Shinn, 2003. Characteristics of word callers:An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. School Psychology Review, 32: 228-240.
3. Klare, G., 1976. Judging readability. Instructional Science, 5: 55-61.
4. O'Connor, R., K. Bell, K. Harty, L. Larkin, S. Sackor and N. Zigmond, 2002. Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. Journal of Educational Psychology, 94: 474–485.
5. Karmakar, Saurav and Ying Zhu, 2010. Visualizing Multiple Text Readability Indexes. Paper presented at 2010 International Conference on Education and Manaement Technology (ICEMT 2010). Cairo.

6. Redish, Janice C. and Selzer Jack, 1985. The Place of Readabilty Formulas in Technical Communication. *Technical Communication*, Fourth Quarter, pp: 1-23.
7. Torres, Marisela and Miguel Roig, 2005. The cloze procedure as a test of plagiarism: The influence of text readability, The Journal of Psychology, 139(3): 221-231.
8. Razek, J.R., G.A. Hosch and D. Pearl, 1982. Readability of accounting textbooks. Journal of Business Education, (October), pp: 23-26.
9. Flory, S.M., T.J. Phillips Jr. and M.F. Tassin, 1992. Measuring readability: a comparison of accounting textbooks, Journal of Accounting Education, 10: 151-161.
10. Plucinski, Kenneth J., 2010 Readability of intermediate Accounting textbooks. Academy of Educational Leadership Journal, 14(2): 49-57.
11. Sabharwal, Sanjeev, Sameer Badarudeen and Shebna Unes Kunju, 2008. Readability of Online Patient Education Materials From the AAOS Web Site, Clinical Orthopaedics and Related Research, 466: 1245-1250.
12. King, Maia M., S.W. Winton Alan and D. Adkins, Angela, 2003. Assessing the Readability of Mental Health Internet Brochures for Children and Adolescents, Journal of Child and Family Studies, 12(1): 91-99.
13. Al-Dawsari, M., 2004. The Assessment of Readability Books Content (Boys-Girls) of the First Grade of Intermediate School According to Readability Standards, Muscat: Sultan Qaboos University.
14. Sinclair, J., 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.
15. McEnery, Tony, 2001. Corpus Linguistics: An Introduction, Edinburgh: Edinburgh University Press.
16. Sulaiti, L. and E. Atwell, 2006. The Design of a Corpus of Contemporary Arabic. International Journal of Corpus Linguistics, 2(7). Amsterdam: John Benjamins.