AARI: Automatic Arabic Readability Index

Abdel-Karim Al-Tamimi[1], Manar Jaradat[2], Nuha Aljarrah[2], and Sahar Ghanim[1]
[1]Computer Engineering Department, Yarmouk University, Jordan
[2]Software Engineering Department, Jordan University of Science and Technology, Jordan

Abstract: *Text readability refers to the ability of the reader to understand and comprehend a given text. In this research, we present our approach to develop an automatic readability index for the Arabic language: Automatic Arabic Readability Index (AARI), using factor analysis. Our results are based on more than 1196 Arabic texts extracted from the Jordanian curriculum in the subjects of: Arabic language, Islamic religion, natural sciences, and national and social education for the elementary classes (first grade through tenth grade). We conduct a comparison study to support our model using cluster analysis and Support Vector Machines (SVM). In order to facilitate the usage of our Arabic readability index, we developed two applications to compute the Arabic text readability: A standalone application and an add-on for Microsoft Word text processer. Through our presented research results and developed tools, we aim to improve the overall readability of Arabic texts, especially those targeted towards the younger generations.*

Keywords: *Readability index, Arabic language, factor analysis, cluster analysis, SVM, text mining.*

Received March 17, 2012; accepted February 21, 2013; published online April 4, 2013

1. Introduction

Text readability refers to the ability of the reader to understand and comprehend a given text. This characteristic depends on many intertwined factors, for instance: the style of writing, the percentage of difficult words contained in the text, the length of the sentences, ..., etc. Readability level is an important indication to determine the possible audiences of a written text and to evaluate the desired impact on its readers. Readability has been widely used in education in order to write and select the appropriate books and assessments for students' level [28]. Its usage is not restricted only to education; it has been widely used in industry for writing manuals and user instructions in a language's level appropriate for the average end-users [3].

Several official agencies require their forms to be written in a manner that meets a specific readability level, in order to better the spread of information among society members, especially among the ones with a lower level of education and a limited literacy. In medicine, the readability of instructions and other important forms, like consent forms, is considered vital to assure better medical treatments and accountability towards patients and their families [22].

With the recent advances in information technology and the widespread of the Internet, readability has been incorporated in search engines and other web tools to correlate web results with the user readability level and to encourage improving the overall writing quality [7, 8]. Additionally, readability index calculations have been incorporated in many text editors like: Microsoft Word, KWord, IBM Lotus Symphony, WordPerfect and WordPro to allow a better utilization of readability scoring [29].

There have been a considerable number of researches concerning readability in languages like English, Spanish and Swedish that resulted in several wide-spread readability indexes [3, 8, 17, 22]. Previous approaches have pursued to achieve a valid readability index using either traditional approaches or using artificial intelligence through methods like Support Vector Machine (SVM) [20]. Traditional approaches are based on using methods like: The reader's judgment or various comprehension questions tests to retrieve a quantifiable indication of the text's readability [13].

The most used approaches in English language revolve around interpreting the lexical complexity features to identify the readability of a certain text. Over 200 mathematical formulas have been published to help assessing the level of text's readability [11]. Some of the most common readability formulas or indexes are: Flesch-Kincaid and Gunning fog indexes for the English language, LIX Index for the Swedish and Danish languages, Fernandez-Huerta index for the Spanish language, and Kandal and Moles index for the French language [7, 28].

Arabic is the first language to more than 280 million and is the second language to 250 million more people [3]. Arabic is also the language of Quran, the holy book of Islamic religion with more than 1.5 billion followers [27]. Despite the fact that Arabic is ranked 5th of the top ten most spoken languages world-wide, only one percent of all blogging content is in Arabic. The number of Arabic blogs has reached 490,000 blogs [1].

Arabic language processing is one of the most active research areas in the region [19]. Despite the urgent need for an Arabic readability index, few researches have been conducted on the Arabic language because of its relative complexity. In [2, 4, 5, 19], the authors researched the problem of Arabic text readability mostly using traditional methods. In addition, most of these researches aimed at measuring the Arabic readability for a specific educational level/ class [4, 19].

In [2], the authors described two Arabic formulas by the names of: Dawood index and Al-Heeti index. Dawood index uses five text lexical features to extract the readability of the Arabic text: average word length, average sentence length, word frequency, percentage of nominal clauses, and the percentage of definite nouns. Al-Heeti index uses only one feature: average word length [5]. The selection of these incorporated features had not been thoroughly justified. This research aims to provide a systematic and scientific approach in determining the main factors that affect the readability of an Arabic text, and to find a mathematical representation of the relationship between the student levels/grades and their corresponding readability index values.

We are in a dire need to measure the readability of the Arabic texts to improve the overall understanding and spread of information written in Arabic. Our research aims to advance the research in Arabic readability and to automate the process of measuring Arabic text readability in a manner similar to other languages.

This paper is divided as follows: Section 2 discusses the main steps performed to insure a proper number of texts is collected and processed correctly. We discuss our early results in performing readability analyses on our collection of Arabic texts. Section 3 explains thoroughly our methodology in determining the principal factors that influence Arabic text readability, and our approach in developing a new Arabic readability formula. Section 4 describes our results obtained using cluster analyses and evaluates our developed index prediction accuracy using SVM. Finally, section 5 concludes our paper and discusses our main contributions.

2. Text Entry and Processing

In this section, we discuss our approach in extracting our Arabic text collections and discuss our early analyses results. Our results are based on more than 1196 Arabic texts extracted from the Jordanian curriculum in the subjects of: Arabic language, Islamic religion, natural sciences, and national and social education for all elementary classes. Since the Jordanian curriculum is not available in an electronic format, we resorted to entering the entire collection of Arabic text manually. In order to facilitate the data-entry operations and other necessary pre and post processing steps, we developed several standalone applications and stored all the extracted texts in a database with the entity-relationship diagram shown in Figure 1. The collection contains more than 1196 texts that consist of more than 432,250 words.

We performed several steps to pre-process the collected data and to prepare it for the succeeding steps of feature extractions and text analyses, as suggested in [2]. The pre-processing process steps include:

1. Removing the punctuation marks, and Arabic diacritical marks.
2. Normalizing the spacing between the words.
3. Converting the Arabic letters {إ , أ and آ to ا}, and { ة to ه}.

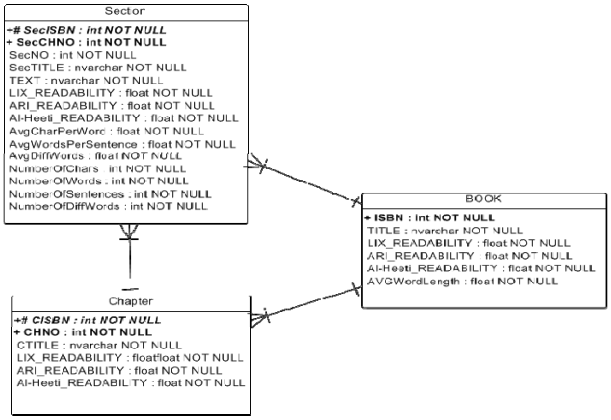

Figure 1. Arabic texts database design.

Feature extraction is the process of extracting all possible features of a text to provide it with representative numerical value(s). As we have mentioned previously, there are many factors associated with text's lexical content and are considered important to determine the readability of any given text. In this step, we extracted seven features that have been used and suggested in several readability formulas [2, 12, 18, 25]:

1. Number of characters in text.
2. Number of words in text.
3. Number of sentences in text.
4. Number of difficult words in text, where our definition of a difficult word is defined as the words consisting of more than six letters after removing " ال" from the beginning of the word, as suggested in [2].
5. Average sentence length: This feature is one of the basic lexical features and is used by most of readability formulas. This feature is calculated as follows:

$$Average\ Sentence\ Length = \frac{Number\ of\ Words\ in\ Text}{Number\ of\ Sentences\ in\ Text} \quad (1)$$

6. Average word length: This feature is calculated as follows:

$$Average\ Word\ Length = \frac{Number\ of\ Letters\ in\ Text}{Number\ of\ Words\ in\ Text} \quad (2)$$

7. Average number of difficult words: This feature is calculated as follows:

$$Avg.\ No.\ of\ Difficult\ Words = \frac{No.\ of\ Difficult\ Words\ in\ Text}{No.\ of\ Words\ in\ Text} \quad (3)$$

As a starting point to identify the proper index to be used for the collected Arabic texts, we compared the following readability indexes:

1. Automated Readability Index (ARI) for English language [25]:

$$ARI\ Grade\ Level = (4.71 \times ACW) + (0.5 \times AWS) - 21.43 \quad (4)$$

Where,
ACW = Average number of chars per word.
AWS = Average number of words per sentence.

2. Lesbarheds Index (LIX) [16]:

$$LIX\ Score = \frac{W}{S} + 100 \times \frac{WD}{W} \quad (5)$$

Where,
W = Number of words.
S = Number of sentences.
WD = Number of difficult words.

3. Al-Heeti readability formula for Arabic language:

$$Al - Heeti\ Grade\ Level = (AWL \times 4.414) - 13.468 \quad (6)$$

Where, *AWL* = Average word length.

To compare our proposed index to other widely used formulas, LIX, ARI and Al-Heeti formulas were selected for their simplicity and more importantly because their parameters can be easily applied to Arabic texts. These indexes are also chosen for the fact that they do not use language-dependent features like number of syllables in a word, as it is the case in Gunning's fog index [7].

Al-Heeti readability formula score indicates the grade level required to comprehend the processed Arabic text. ARI readability formula score refers to the US grade level needed to comprehend the text. LIX readability formula produces a score that determines the difficulty of a given text. Table 1 lists LIX readability scores and their meanings.

Table 1. LIX scores and their meanings.

Score	Meaning
0-24	Very easy
25-43	Easy
35-44	Standard
45-54	Difficult
55 and above	Very difficult

We developed a standalone application to automate the calculation of the three readability indexes values for a given Arabic text, and facilitate the comparison process. The application performs all the necessary pre-processing steps mentioned previously. We have computed the main seven features previously mentioned and calculated the three readability indexes (Al-Heeti, LIX, and ARI) for our entire collection of Arabic texts.

The results of our statistical calculations are summarized in Tables 2, 3, and 4. As Table 2 shows, the Average Word Length (AWL) feature used in Al-Heeti is not a good indication of the readability level of an Arabic text, as all the grades have similar values and thus one cannot distinguish and specify the readability level solely on the average word length. This conclusion is further confirmed in the results obtained in Table 4, as will be discussed shortly. We can notice in Table 2, only the average words per sentence values give a good indication of the readability level progress among grade levels.

Table 2. Main text statistics for Arabic text collection.

Grade Level	Average Word Length	Average Words Per Sentence	Average Difficult Words
1st Grade	4.33	6.45	0.038
2nd Grade	4.26	8.41	0.036
3rd Grade	4.77	10.21	0.075
4th Grade	4.25	16.46	0.045
5th Grade	3.93	17.81	0.035
6th Grade	4.16	23.02	0.048
7th Grade	4.17	12.19	0.046
8th Grade	4.26	33.68	0.066
9th Grade	4.29	19.86	0.055
10th Grade	4.17	27.46	0.056

In Table 3, the results are not as easy to interpret. There is no obvious trend in both Number of words and Number of sentences data columns. The presence of trends is important to determine the most important parameters that affect and represent the text's Arabic readability.

Table 3. Main text statistics for Arabic text collection.

Grade Level	Number of Characters	Number of Words	Number of Sentences	Number of Difficult Words
1st Grade	51076	12005	2066	460
2nd Grade	65092	15592	2344	513
3rd Grade	31608	6752	747	468
4th Grade	194976	46021	3453	2059
5th Grade	245661	62975	4066	2131
6th Grade	210729	50714	2821	2561
7th Grade	246207	58318	5825	2831
8th Grade	120469	29042	979	1658
9th Grade	262495	60753	4225	3420
10th Grade	262844	63236	2903	3446

We have also computed LIX, ARI, and Al-Heeti readability formula values for all the grades texts as shown in Table 4. As we can notice, only LIX score provides near consistent results in indicating the readability of an Arabic text. Alas, since LIX was not developed nor optimized to be used with Arabic texts, the LIX results paint all processed Arabic texts as very-easy as previously indicated in Table 1. Both ARI and AL-Heeti give inconsistent and unreliable results. In addition, similar to the conclusion point discussed in

[2], we have noticed that ARI gives unacceptable negative score values for some Arabic texts.

Table 4. Average readability scores for Arabic text collection.

Grade Level	LIX Score	ARI Grade Level	Al-Heeti Grade Level
1st Grade	10.31759	2.230111	5.682323
2nd Grade	12.0681	2.852769	5.343665
3rd Grade	17.78686	6.175685	7.615622
4th Grade	21.03279	6.839132	5.30756
5th Grade	21.36831	5.986994	3.880371
6th Grade	27.87063	9.708773	4.922915
7th Grade	16.86823	4.354215	4.981254
8th Grade	40.30262	15.51823	5.374064
9th Grade	25.45378	8.740915	5.49867
10th Grade	33.09925	11.97561	4.966595

Due to the ambiguity of the relationship between the extracted text lexical features and the processed Arabic text, in addition to the inaccuracy and inapplicability of other tested readability indexes on Arabic texts, the need is emphasized to consider a new Arabic readability index using factor analysis. In the next section, we will demonstrate our research process and findings in determining the most influential text features using factor analysis.

3. Readability Factors and Factor Analysis

As mentioned before, the readability of a text depends on the features that are related to the reader as well as the text itself. The reader features include: their language capability, their background knowledge of the subject matter, and their motivation to read the text. In addition, there are physical features which may affect the text's readability including: the font size, the design clarity, the layout and extra textual features like pictures and diagrams.

The most important set of factors that affect readability are the factors that are related to the text itself, which include: word length, word frequency, vocabulary load, number of difficult words, average sentence length, sentence complexity, the clarity of the text idea, the use of topology or metaphors, and the grammatical structure complexity. In order to develop a valid readability formula, all these factors should be taken into consideration.

As factors vary, not all factors have the same impact on the readability measurement. Each factor has its own weight or load in affecting the text readability. Such that, some factors have a small effect on the readability measure compared to other factors, and thus play a more important role in determining the text readability.

In this section, we discuss the principal steps we performed using several factor analysis techniques in order to find the main factors that determine the readability of a given Arabic text. These factors are then incorporated to form our new Arabic readability formula.

The main goal of the factor analysis phase is to reduce the number of redundant factors and to detect the structure of the relationships among the considered variables. Factor analysis as a reduction method simplifies complex multivariate dataset by finding a natural grouping within the data. In such a natural grouping, all variables within a particular group are highly correlated among themselves but have a small correlation with the variables in other groups. Thus, each group of variables can be represented by a single factor. Table 5 shows the correlation between the seven text readability features. As we can notice from the table, there is a high correlation among different text features, as shown in the relationship between number of characters and number of words. In order to group the possibly correlated features, we applied Principal Component Analysis (PCA) method. Our aim is to provide a smaller set of uncorrelated factors that represent most of the variance of the entire set of factors. We have performed our factor analysis steps using R statistical software [23].

The importance of each variable is represented by its eigenvalue. We used Kaiser-Guttman rule [15] to determine the number of uncorrelated factors to be extracted. Based on this rule, we excluded the parameters with eigenvalues less than one. As Table 6 shows, only the parameters 1, 2, and 3 have an eigenvalue greater than one.

These three factors represent 91.97%, [(3.488 + 1.756 + 1.194)/ 7] of the total standardized variance. To support our selection, we performed the Scree test [9], where we plot the number of text features (candidate factors) and the total variance of the Arabic text collection. The results based on the Scree test suggest the use of three factors. Additionally, as shown in Figure 2, we performed several non-graphical tests to confirm that the number of factors is sufficient to explain the inter-correlation among the selected features [24].

Table 5. Correlation between selected text readability features.

Variables	Avg. Chars Per Word	Avg. Words Per Sentence	Avg. of Difficult Words	No. of Characters	No. of Words	No. of Sentences	No. of Difficult Words
Avg. Chars Per Word	1.000	-0.222	0.649	-0.027	-0.122	-0.023	0.234
Avg. Words Per Sentence	-0.222	1.000	0.047	0.319	0.359	-0.221	0.234
Avg. of Difficult Words	0.649	0.047	1.000	0.129	0.071	0.034	0.428
No. of Chars	-0.027	0.319	0.129	1.000	0.991	0.727	0.868
No. of Words	-0.122	0.359	0.071	0.991	1.000	0.710	0.816
No. of Sentences	-0.023	-0.221	0.034	0.727	0.710	1.000	0.618
No. of Difficult Words	0.234	0.234	0.428	0.868	0.816	0.618	1.000

Table 6. Text features eigenvalues.

Variables		Eigenvalue
Average character per word	(1)	3.488
Average word per sentence	(2)	1.756
Average of difficult word	(3)	1.194
Number of characters	(4)	0.308
Number of words	(5)	0.159
Number of sentences	(6)	0.092
Number of difficult words	(7)	0.002

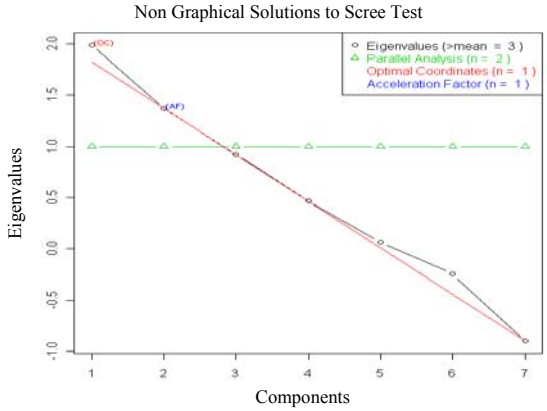

Figure 2. Number of principal components using non-graphical solution to Scree test.

In order to spread out the correlation between variables and simplify the factors structure, we performed both orthogonal (varimax) and non-orthogonal (oblique) rotation methods [16]. Table 7 shows the loading values for both un-rotated and rotated (varimax, oblique) factors. It is clear from the results obtained that the un-rotated factors distribution hides a grouping in data. Thus, we chose the orthogonal (varimax) rotation method since it is the most effective in spreading out the factors to make them simpler to interpret and group. We can notice that the oblique method results are similar, and thus will convey the same factor analysis results.

Based on Table 7, we can note that the First Factor (VF1) defines the number of characters, and number of words features. The Second Factor (VF2) represents average number of characters per word. The Third Factor (VF3) defines the average words per sentence feature. Figure 3 shows the three main groups distribution in the space of principal components.

We chose the number of characters feature to represent the VF1, since it has the highest load (0.98). For VF2, we chose the average characters per word (0.80) since it has the highest load and has the least correlation with the number of characters feature (-0.027). VF3 is represented by the average word per sentence feature (0.99). This analyses show the importance of the number of characters, the average characters per word and the average words per-sentence features in determining the readability of an Arabic text. They also show that formulas like Al-Heeti that relies only on one factor are unsuitable for complex languages like Arabic language, as was discussed earlier in the results shown in Table 4.

Based on these results, we can develop a new readability formula to measure the readability of a given Arabic text. Table 8 shows the loading values for each factor, these factors represent the coefficients of each variant: (VF1), (VF2) and (VF3) in our readability formula.

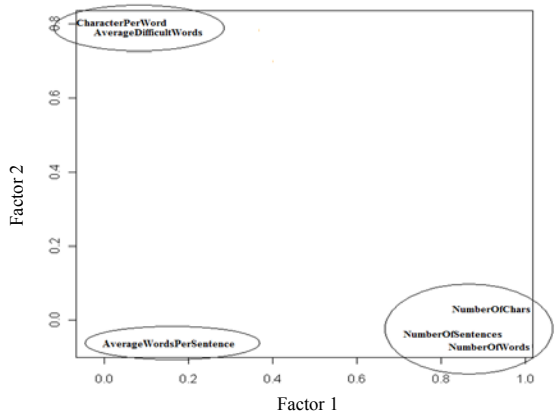

Figure 3. Scatter-plot of varimax rotated factors (OF1) and (OF2) in the space of the principal components.

To represent our results in a simple mathematical formula, we introduce Automatic Arabic Readability Index Base (AARI Base):

$$AARI\ Base = (3.28 \times NOC) + (1.43 \times ACW) + (1.24 \times AWS) \quad (7)$$

Where,
NOC = Number of characters.
ACW = Average character per word.
AWS = Average words per sentence.

Figure 4 shows the standalone application developed to calculate the different readability indexes score and the AARI Base values.

Table 7. Estimated and rotated factor loadings.

Variable	Un-Rotated			Rotated (Varimax)			Rotated (Oblique)		
	F1	F2	F3	VF1	VF2	VF3	OF1	OF2	OF3
Number of characters	0.97	-	-	**0.98**	-	-	0.98	-	-
Number of words	0.98	-	-	**0.97**	-	-	0.98	-	-
Number of sentences	0.81	0.42	-	0.81	-	-0.33	0.83	-	-0.4
Number of difficult words	-	0.79	-	0.84	0.41	-	0.81	0.36	-
Average characters per word	-	0.78	-	-	**0.80**	-	-	0.81	-
Average of difficult word	0.54	-	-0.84	-	0.78	-	-	0.78	-
Average words per sentence	0.59	-	0.65	-	-	**0.99**	-	-	0.98

Table 8. Arabic readability factors loadings.

Factor	Factor Loadings
VF1	3.28
VF2	1.43
VF3	1.24

Our next step is to formulate a relationship between AARI Base value and different grade levels using simple regression techniques. We started with plotting the mean AARI Base index for the ten grades texts, and then we proceeded with removing the AARI Base average outliers. Figure 5 shows the resulted relationship between grade levels and the mean AARI Base values.

Figure 4. Standalone application with Al-Heeti, ARI, LIX, and AARI indexes calculation.

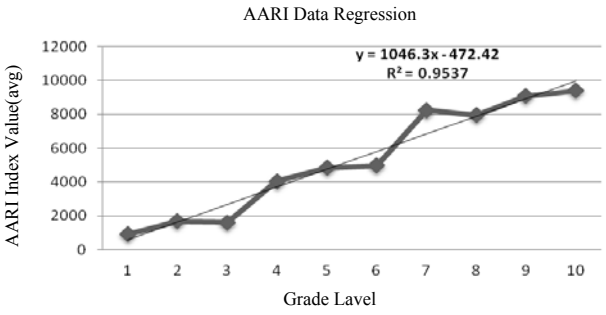

Figure 5. Regression analysis of AARI values.

In the regression analysis step, we experimented with different regression methods, including: linear, *n*-polynomial, exponential and power regression. Simple linear regression achieved a high correlation value between AARI Base average values and the regression line. The correlation value achieved is 95.37%. The resulted regression line equation can be written as the AARI:

$$AARI = 1046.3 \times (Grade\ Level) - 472.42 \quad (8)$$

Thus, the relationship between text's grade level and AARI can be written as:

$$Grade\ Level = \frac{(AARI + 472.42)}{1046.3} \quad (9)$$

We believe that the addition of a readability tool to a popular and wide-spread text processor like Microsoft Word will further facilitate the usage of Arabic readability indexes and increase the exposure of their usefulness in writing professional and educational Arabic texts. Figure 6 shows the implementation of our Microsoft Word add-on. The add-on calculates the readability indexes for AARI, Al-Heeti, LIX and ARI indexes.

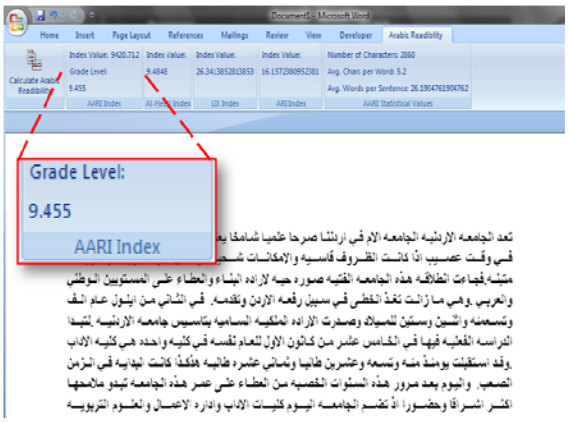

Figure 6. Microsoft word add-on test on a 9th grade sample text.

Table 9 shows the average readability scores using AARI for our Arabic text collection. It is clear that there is a positive correlation between grade levels and average readability scores, which supports the validity of our readability index.

Table 9. AARI base readability scores.

Grade Level	Automatic Arabic Readability Base
1st Grade	924.68
2nd Grade	1684.52
3rd Grade	1622.80
4th Grade	4048.65
5th Grade	4852.65
6th Grade	4971.59
7th Grade	8261.49
8th Grade	7950.64
9th Grade	9093.762
10th Grade	9410.994

4. SVM and Cluster Analyses

In this section, we extend our research analysis to investigate clustering grades (1st grade, 2nd grade, ..., etc.,) into groups to increase the prediction accuracy of AARI. To identify the number of clusters that we need to work with in order to achieve a higher prediction accuracy, we started by performing unsupervised clustering techniques. Which determines the number of clusters into which the data can be grouped. We used one the most common clustering methods: k-means clustering algorithm [21].

PCA gives an insight of how many clusters our grades can be grouped into [10]. In our case, PCA suggests that we need three clusters, based on the number of principle factors, as discussed in section 3.

In order to verify this suggestion, we proceeded with applying k-means clustering by computing the within-cluster sum of squares of the AARI values for a different number of clusters. Our aim is to select the minimum number of clusters that allow the minimal possible value for the within-cluster sum of squares of the AARI values.

By plotting these values, we are presented with a graph similar to the Scree test, where the large spaces between the plotted variables and the graph possible knees indicates that that number of possible clusters is to be either three or four clusters as shown in Figure 7. We decided on using three clusters since the number represents the knee of the Scree test curve better than the other options, and since it is supported by the previous PCA analysis.

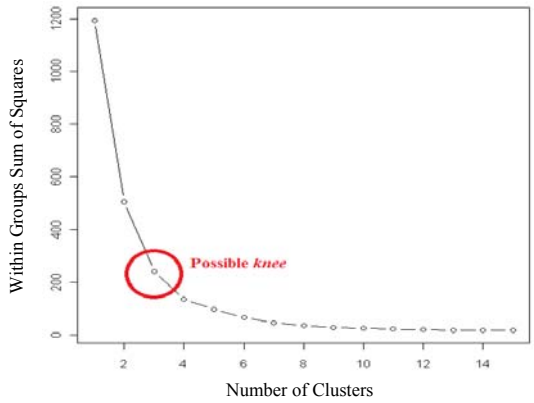

Figure 7. Scree test to determine the number of clusters.

Our next step is to cluster the grades into three clusters using a supervised clustering method. In [2, 3, 20] the authors used SVM to classify texts into clusters depending on their readability level. In our analysis, we used package (e1071) in R to apply SVM classification on our data and classify them depending on their average AARI values. We proceeded by dividing our text collection into two parts: 70% used as a training set, and 30% is used as validation set.

We performed a C-SVM classification on our training-set using the radial basis function kernel performed on their AARI values [26]. This resulted in three clusters divided as shown in Table 10. These results confirm the visual identification of clusters that can be performed on Figure 7.

Table 10. Grades clustering using C-SVM.

Cluster	Grade Levels
Cluster 1	1st grade, 2nd grade, and 3rd grade
Cluster 2	4th grade, 5th grade, and 6th grade
Cluster 3	7th grade, 8th grade, 9th grade, and 10th grade

After that, we conducted several prediction tests using the remaining 30% of our text collection. We have achieved an accuracy rate in our prediction test that reached 83.23% on average for the ten grades over ten consecutive runs (Figure 8).

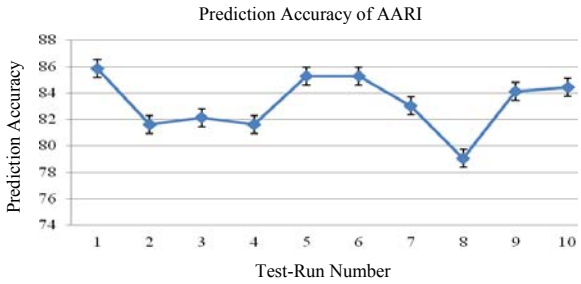

Figure 8. Prediction accuracy of AARI over ten test-runs.

To prove the validity and importance of our clustering analysis, we repeated our prediction accuracy tests by dividing the data into ten clusters instead. Each cluster contains only one grade level, simulating no-clustering configuration. The prediction results reached only 49.32% level of accuracy on average. Such results confirm that clustering grades on the right criterion, in this case AARI values, results in better predictability rates. In the next section, we summarize the contributions provided in this paper and discusses the importance of our findings.

5. Conclusions

In this paper, we presented our approach to develop an accurate and easy to implement AARI. We conducted our analyses using our collection of Arabic texts with more than 1190 texts extracted from the Jordanian curriculum. We demonstrated the shortcomings of the available readability formulas, and their ineffectiveness to produce accurate readability scores for Arabic texts.

We have also determined the main factors that influence the readability of Arabic texts using factor analysis. We have showed that some of the lexical features of an Arabic text out-weigh other factors that were commonly used in other readability formulas.

AARI, as we have discussed in this paper, provides an accurate representation of Arabic text readability. AARI can be represented using a simple mathematical equation that allows it to be incorporated easily in any readability tool. We extended our research analyses to investigate the possibility of clustering grades level using k-means clustering method to improve its readability index prediction accuracy. Subsequently, we tested our clustering methodology using C-SVM classification. We concluded that our classification approach achieves up to 83.23% in prediction accuracy, if no clustering is used, the accuracy drops to 49.32%.

We aim through our contributions to advance the research in Arabic text readability field, and we anticipate that through the usage of our developed readability index tools and developed readability index (AARI) we can help improve the spread of Arabic texts over the web, and improve the quality of Arabic texts in general.

References

[1] Al-Ajlan A., Al-Khalifa H., and Al-Salman A., "Towards the Development of an Automatic Readability Measurements for Arabic Language," *in Proceedings of the 3rd International Conference on Digital Information Management*, London, United Kingdom, pp. 506-511, 2008.

[2] Al-Dawsari M., "The Assessment of Readability Books Content (Boys-Girls) of the First Grade of Intermediate School According to Readability Standards," *Technical Report*, Sultan Qaboos University, Oman, 2004.

[3] Al-Khalifa H. and Al-Ajlan M., "Automatic Readability Measurements of the Arabic Text: An Exploratory Study," *the Arabian Journal for Science and Engineering*, vol. 35, no. 2C, pp. 103-124, 2011.

[4] Al-Naji H., "Readability Level of the Reading Book of the 6th Grade of Elementary School in United Arab Emirates," available at: http://www.arabic1.org.sa, last visited 2011.

[5] Al-Talhi A., "Readability Level Measurement of High Elementary Students in Makkah, Taif and Jeddah," *Technical Report*, Umm Al-Qura University, Makah, Saudi Arabia, 1994.

[6] Araj A., "The Relationship between the Mental Skills Reflected by the Readability Test of Arabic Texts and the Achievement Level of Arabic Language for the Fifth Grade in Bethlehem District," *Master Thesis*, Faculty of Education, Al-Quds University, Palestine, 1999.

[7] Aula A., "Enhancing the Readability of Search Result Summaries," *in Proceedings of HCI Conference, Design for Life*, Leeds, UK, vol. 2, pp. 1-4, 2004.

[8] Basu S., "8 Readability Web Tools to Test Your Writing Quality," available at: http://www.makeuseof.com/ tag/writing-reader-friendly-check-8- readability- testing-web-tools/, last visited 2011.

[9] Cattel R., "The Scree Test for the Number of Factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245-276, 1966.

[10] Ding C. and He X., "K-Means Clustering via Principal Component Analysis," *in Proceedings of the 21st International Conference on Machine Learning*, New York, USA, vol. 69, pp. 29, 2004.

[11] DuBay W., "The Principles of Readability," available at: http://files.eric.ed.gov/fulltext/ ED490073.pdf, last visited 2004.

[12] Gunning R., "The Fog Index After Twenty Years," *Journal of Business Communication*, vol. 6, no. 2, pp. 3-13, 1969.

[13] Hall R., "Islamic Spirituality Vis-a-Vis Asia Pacific Muslim Populations: A Resource for Western Social Work Practice," *International Social Work*, vol. 55, no. 1, pp. 109-124, 2011.

[14] Haraty R. and Ghaddar C., "Arabic Text Recognition," *the International Arab Journal of Information Technology*, vol. 1, no. 2, pp. 156-163, 2004.

[15] Kaiser H., "The Application of Electronic Computers to Factor Analysis," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 141-151, 1960.

[16] Kootstra G., "Project on Exploratory Factor Analysis Applied to Foreign Language Learning," 2004.

[17] Krantz P., "Methods for Measuring Text Readability," available at: http://www.languages.ufl.edu/ languages.html, last visited 2011.

[18] Lau T., *Chinese Readability Analysis and its Applications on the Internet*, Chinese University of Hong Kong, Hong Kong, 2006.

[19] Liu X., Croft W., Oh P., and Hart D., "Automatic Recognition of Reading Levels from User Queries," *in Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, USA, pp. 548-549, 2004.

[20] Liu X., Croft W., Oh P., and Hart D., "Automatic Recognition of Reading Levels from User Queries," *in Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, United Kingdom, pp. 548-549, 2004.

[21] Norusis M., *SPSS 17.0 Statistical Procedures Companion*, Prentice Hall, USA, 2009.

[22] Pasche-Orlow M., Taylor H., and Brancti F., "Readability Standards for Informed-Consent Forms as Compared with Actual Readability," *the New England Journal of Medicine*, vol. 380, no. 2003, pp. 721-726, 2003.

[23] R-Project for Statistical Computing, available at: http://www.r-project.org, last visited 2013.

[24] Raîche G., Riopel M., and Blais J., "Non Graphical Solutions for the Cattell's Scree Test," *Journal of Research Methods for the Behavioral and Social Sciences*, Montréal, Canada, vol. 9, no. 1, pp. 23-29, 2013.

[25] Si L. and Callan J., "A Statistical Model for Scientific Readability," *in Proceedings of the 10th International Conference on Information and Knowledge Management*, New York, USA, pp. 574-576, 2001.

[26] StatSoft, Support Vector Machines, available at: http://www.statsoft.com/ textbook/support-vector-machines/, last visited 2011.

[27] The Initiative for an Open Arab Internet, "Arabic Blogs," available at: http://old.openarab.net/ en/node/1638, last visited 2011.

[28] Ulusoy M., "Readability Approaches: Implications for Turkey," *International Education Journal*, vol. 7, no. 3, pp. 323-332, 2006.

[29] Wikipedia Article, "Flesch-Kincaid Readability Test," available at: http://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_test, last visited 2012.

Abdel-Karim Al-Tamimi is an assistant professor in the Computer Engineering Department at Yarmouk University, Jordan. He received his BSc degree in computer engineering from Yarmouk University. Then, he received his Master and PhD degrees in computer engineering from Washington University in St. Louis in 2007 and 2010, respectively. His research interests include computer networks, multimedia networks and applications, modeling and simulation, and computer security.

Manar Jaradat received her BSc degree in computer engineering from Yarmouk University, Jordan. Currently, she is working as a lab instructor at the Software Engineering Department in the University of Science and Technology, Jordan.

Nuha Aljarrah received her BSc degree in computer engineering from Yarmouk University, Jordan. Currently, she is working at the Network Engineering and Security Department in the University of Science and Technology, Jordan.

Sahar Ghanim received her BSc degree in computer engineering from Yarmouk University, Jordan.