A Pearson Product-moment Correlation Coefficient Based Analysis of Comparison of Character and Syllable-based Readability Indices of Web-pages

Dr. Jatinderkumar R. Saini

Abstract— With an increased exposure of internet, usage of matrimonial sites has increased. The providers of matrimonial services often require the users to sign-up and agree to their terms and conditions of usage. These conditions differently referred to by matrimonial web-sites as private policy, safety guide, etc. have been here collectively identified as policy guides. This paper presents the analysis of comparison of two sets of readability indices for more than 50 webpages comprising nearly 90,000 words from 30 matrimonial websites. The first set of readability indices is composed of two character-based tests viz. Automated Readability Index (ARI) and Coleman-Liau Index (CLI) whereas the second set is composed of three syllable-based tests viz. Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI) and Simple Measure of Gobbledygook (SMOG) Index. The averages of the two sets have been found to be 11.97 and 11.18, respectively whereas the Pearson Product-moment Correlation Coefficient value of 0.76 has been found. Both the results lead to conclude that the set of character-based and the set of syllable-based readability indices could be used interchangeably for assessing the comprehensibility of a text document. The paper also presents the results of inter-set and intra-set comparison of individual indices from both sets.

***Index Terms*—** Matrimonial Websites, Pearson Product-moment Correlation Coefficient, Privacy Policy, Readability Formula, Terms and Conditions

I. INTRODUCTION

As more and more people rely on the wealth of information available online, there has been an increase in the exposure on the internet. The Internet user population was more than one billion persons by the end of year 2009 [1]. The same figure for the first quarter of year 2010 was 1.83 billion with the erstwhile estimated projection of 2.10 billion for the year 2012 [2]. According to Gyongyi and Garcia-Molina [3], most frequently, search engines are the entryways to the Web. Additionally, email has been an efficient and popular means of electronic communication. But search engines and emails are just two of the many doorways of internet. The other access paths of the online world include social networks, dating sites, online messengers, blogs, matrimonial sites and the sites offering services like hotel booking, flight booking and railway reservation, to name a few.

Most often, the service provided by a particular web-site is governed and controlled by the terms of usage of web-site. These terms of usage are provided in various forms like 'private policy', 'terms and conditions of usage' and 'safety guide'. This paper refers to them collectively as policy guides. These policy guides are like documents of 'dos' and 'donts' intended for the users of the website. In many cases, the web-site requires the user to sign-up for being able to use the service provided by the web-site. This signup also requires the user to agree to the conditions presented by the web-site and only user's agreement to such conditions leads to a successful sign-up. Once, the user has signed-up successfully, the same information could be subsequently used by the user for sign-in purpose. Even though it is important, generally quite a few people bother to go through and read the full terms and conditions of agreement. But the other section of people who does read the policy guides, need also to understand its contents.

Readability statistics are defined by RFP Evaluation Centers [4] as indicators, under the form of readability scores, which measure how easily an adult can read and understand a text. Readability statistics are therefore a good predictor of the level of difficulty of particularly technical documents. Readability statistics present different readability scores that are computed using readability formulas. According to RFP Evaluation Centers [4], the most commonly used readability statistics formulas are:

- Passive Sentences
- Flesch Reading Ease (FRE)
- Flesch-Kincaid Grade Level (FKGL)
- Coleman-Liau Grade Level (CLGL)

Manuscript received Aug 20, 2014.

Dr. Jatinderkumar R. Saini, Associate Professor & I/C Director, Narmada College of Computer Application, Bharuch, Gujarat, India

- Bormuth Grade Level (BGL)

Coleman-Liau Grade Level (CLGL) is also called Coleman-Liau Index (CLI). In addition to CLI, Percentage of Passive Sentences (PPS), Gunning Fog Index (GFI), Automated Readability Index (ARI) and Simple Measure of Gobbledygook (SMOG) Index are also important indices for estimation of readability of text. Readability scores assess the reading level of a document. RFP Evaluation Centers [4] has further provided a descriptive note on each of these readability statistics formulas. According to them, Passive Sentences readability statistics formula provides the ratio of passive sentences over active sentences. The FRE readability statistics formula rates text on a 100-point scale based on the average number of syllables or characters per word and words per sentence. The higher the score, the easier it is to understand the document. The FKGL readability statistics formula rates text on a United States (U.S.) grade-school level, e.g. a score of 8.0 means that an eighth grader can understand the document. ARI, CLI, GFI and SMOG Index are interpreted similar to FKGL. The formulas for the indexes under consideration of present paper, as provided by editcentral.com [5] are presented in Table I. Even though these indices work with similar inputs but there results generally do not coincide, owing to the way of calculations and constants involved in the respective formulas. The consideration of characters instead of syllables and vice versa also influences the results.

Table I: The Formulas of Readability Scores

Sr. No.	Readability Score	Formula
1	Automated Readability Index (ARI)	4.71 x (c/sw) + 0.5 x (sw/s) – 21.43
2	Coleman-Liau Index (CLI)	(0.0588 x L) – (0.296 x S) – 15.8
3	Flesch-Kincaid Grade Level (FKGL)	(0.39 x ASL) + (11.8 x ASW) – 15.59
4	Gunning Fog Index (GFI)	0.4 x [(w/s) + 100 x (cw/w)]
5	Simple Measure of Gobbledygook (SMOG)	sqrt(cw x 30.0/S) + 3.0

where:

w (for words) is the number of words in a document

s (for sentences) is the number of sentences in a document

cw (for complex words) is the number of complex words that is those with three or more syllables

L (for letters, average value) is the average number of letters per 100 words

S (for sentences, average value) is the average number of sentences per 100 words.

c (for characters) is the number of letters, numbers, and punctuation marks

sw (for space-words) is the number of spaces

ASL (for average sentence length) is the number of words divided by the number of sentences

ASW (for average number of syllables per word) is the number of syllables divided by the number of words

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name [6]. It is calculated by taking the ratio of covariance of two variables by the product of the standard deviations of the two variables.

II. Literature Review

Meade et. al. [7] designed a study to determine if simplification of smoking literature improved patient comprehension of the literature. They divided patients under study into different groups and provided each group with a literature written at different school grade level. Their findings were that those receiving lower grade literatures showed better comprehension than those receiving higher grade literatures. Based on this, they have concluded that comprehension of written smoking materials can be improved by adjustment of the reading grade level. Another important finding of Meade et. al. [7] is that educational level is a poor predictor of reading ability. This means to say that years of schooling do not identify reading ability but simplicity of the text does. McNamara et. al. [8] in their research work aimed at predicting text readability and facilitating text comprehension. They argue that there is a pressing need to improve reading comprehension of the text for its better grasping.

From the perspective of Web Content Mining and Text Mining, the related literature includes works related to analysis of spam emails [9], textual analysis [10], structural analysis [11] and character-usage analysis [12] of email addresses. The author of the current paper believes that interest in reading a given text is also plays an important role in the comprehension of the text. This means to say that the probability of comprehension of interesting text is more than the probability of comprehension of text of no interest to the reader. The readability formulas judge the appropriateness of text based on their ease or difficulty of comprehension. In this context, the author agrees with Belloni and Jongsma [13] who are of the opinion that there is a dire effect of interest on reading

comprehension. This is more so in context of presented work because the interest of readers in going through the policy guides is very less.

Saini [14] had worked on analyzing the comprehension ease of policy guides of dating sites. His experiments have shown that the average value for PPS was 3% more than the expected standard value; the average value for FRE was approximately 50% of the expected standard value while the average value of FKGL was 5 points more than the expected standard value. The deviations from standard values were all in directions contributing to low readability of policy guides of dating sites. In other similar works, the understandability of policy guides of matrimonial websites using metric scores of readability formulas was evaluated using PPS, FRE and FKGL as the basic metrics [15]. He found that the average value for PPS was 5% more than the expected standard value; the average value for FRE was approximately 25% lower than the expected standard value while the average value of FKGL was 3 points more than the expected standard value. The deviations from standard values were all in directions contributing to low readability of policy guides of matrimonial websites too. When GFI, CLI and ARI were used as basic metrics [16], it was found that the average value for GFI and CLI is more than the expected standard value. Both these deviations from standard values are in direction contributing to low readability of policy guides of matrimonial websites. The average value of ARI was found within the range of standard expected value. The overall readability of policy guides of matrimonial websites was found to be low since the average values of GFI, CLI and ARI scores averaged to a value of 12.22 which was more than the expected standard value of 11.00.

Mateus [17] had discussed the improvement of web-site readability based on factors like formatting features of font and content density on the web-page. He has also elaborated on the usage of background and foreground color for improving the readability of the web-page. Sedlak [18] has also elaborated on the careful use of aspects like contrasting colors, bullet points and subheadings for good web design and increasing the readability of the web-site.

Through the survey of related works, it has been found that the readability of various kinds of text has been studied. The previous researchers have also worked towards the improvement of readability through the employment of various kinds of readability scores. Attempts have also been made on issues related to the readability of web-pages as also on effects of readability on comprehension of web-pages. This paper elaborates on analysis of comparison of character-based and syllable-based readability indices, using the Person Product-moment Correlation Coefficient for policy guide web-pages of matrimonial web-sites.

III. Methodology

This section presents the strategy adopted for comparison of character and syllable-based readability indices of webpages. In order to populate the text corpus, the web text was decided to be obtained from similar kind of websites. This was done to assure consistency in the readability indices of webpages with a consideration that the webpages of similar kind of websites will have inter-website consistent readability indices. The matrimonial websites were chosen for this purpose. From social perspective, the author agrees that the matrimonial websites or marriage websites are only a variation of the standard dating websites [19]. From the technical perspective the analysis of textual content of webpages is a Web Text Content Mining task and deals with Natural Language Processing (NLP).

The process started with the collection of a list of 49 sites offering matrimonial services. For this collection, the care was taken that the site is specifically providing the matrimonial service and is not dedicated to any service like dating, pornography and social networking. This was done to assure that the web-page to be dealt with is solely dedicated to matrimonial service and hence is more specific to matrimony instead of being of general nature for similar web-pages. Out of the collected list of 49 sites providing matrimonial services, 5 sites were further discarded because they were just providing the directory of other matrimonial sites and were therefore of no direct relevance to the objective of this research. Definitely, this directory was used to explore other matrimonial sites to get their policy guides. Further, another 4 sites were discarded as they did not allow fetching the textual portion of their policy guides. Another 4 sites have to be discarded as their website Uniform Resource Locator (URL) was pointing/re-directing to a site which had already been considered. Finally, last set of 4 websites have to be discarded because they did not had policy guides. Thus, out of a collection of 49 websites, a total of 17 have to be discarded.

From the available list of 32 matrimonial sites, a set of web-pages dealing with policies of sites was formed. The set of web-pages thus created consisted of web-pages dealing with Terms and Conditions of the usage of the respective site. During this process, it was found that there are many similar terminologies used by web-sites for providing their usage-policy. Though technically different, these web-pages have been considered as of similar nature and treated all as usage-policy guides of the matrimonial web-sites. The matrimonial web-sites mainly provide their policy guides in the following forms and variations:

- Confidentiality Features
- Disclaimer / Disclaimer Agreement / Disclaimer Terms and Conditions
- Privacy Guide/Policy/Statement
- Safety Guide/Tips
- Service Agreement
- Terms and Conditions & Policy Guide
- Terms and Conditions Guide
- Terms of Use
- Usage Agreement

For simplicity, the current paper has normalized these phrases and only the unique, distinct and superset phrases for the above set have been considered. An instance of normalization is considering the phrase 'Terms and Conditions' for occurrences like 'Terms of Use', 'Terms', 'Conditions of Use' and 'Terms and Conditions Agreement'. Following is the list of normalized phrases:

- Disclaimer
- Privacy Policy
- Safety Guidelines
- Terms and Conditions

It was further found that some matrimonial web-sites provide all of their policy terms and conditions through a single policy guide while others provide them through multiple and different web-pages. Accordingly, the 32 web-sites yielded a total of 58 web-pages of which each provided policy guidelines regarding one of the safety usage, general terms and conditions and privacy statement. The objective was to analyze these 58 web-pages on the basis of their readability scores. A notable thing here is that the web-pages of a 'multiple web-page policy guide' of a web-site were not merged into a single text corpus. This was done to prevent the dilution of readability score of each web-page. Hence, the text of each web-page of policy guide was copied to a separate text document.

The corpus of text created from matrimonial websites was sourced in such a way that it constitutes a mixed input from matrimonial websites of different regions and religions. For better computation of readability scores, the minimum number of words required in a document is 200 [4]. It was found that three of the 58 web-pages were having 103, 52 and 191 words only. Consequently, they were removed from the set of web-pages forming collection of policy guides. This also resulted in reduction in number of two policy guides, one having one policy page while the other having two policy pages. Subsequently, the minimum number of words in remaining 55 web-pages of 30 websites yielding 30 policy guides was 222. The number of words for each of the policy pages is presented in a tabular form in Table II.

These web-pages were subjected to the calculation of the following readability indices for obtaining their statistical values:

- Automated Readability Index (ARI)
- Coleman-Liau Index (CLI)
- Flesch-Kincaid Grade Level (FKGL)
- Gunning Fog Index (GFI)
- Simple Measure of Gobbledygook (SMOG) Index

ARI and CLI are character-based readability indices while FKGL, GFI and SMOGI are syllable-based readability indices. All of these five readability indices were calculated along with the averages of two character-based readability indices as well as the average of three character-based readability indices.

Table II: Title of Policy Web-page and Corresponding Number of Words for Matrimonial Sites

Policy Guide Number	Abbreviated Policy Guide Number (APG#)	Title of Web-page (Policy Page No.)	Abbreviated Policy Page No. (APP#)	No. of Words
Policy Guide 1	PG1	Terms and Conditions	PP1	227
Policy Guide 2	PG2	Privacy Policy	PP2	448
		Terms and Conditions	PP3	1673
Policy Guide 3	PG3	Privacy Policy	PP4	462
		Terms and Conditions	PP5	4090
Policy Guide 4	PG4	Terms and Conditions	PP6	1915
Policy Guide 5	PG5	Privacy Policy	PP7	1249
Policy Guide 6	PG6	Disclaimer	PP8	623
		Privacy Policy	PP9	1617
		Safety Guidelines	PP10	785
		Terms and Conditions	PP11	2897
Policy Guide 7	PG7	Terms and Conditions	PP12	5031
		Privacy Policy	PP13	453
Policy Guide 8	PG8	Privacy Policy	PP14	1305
Policy Guide 9	PG9	Terms and Conditions	PP15	2877
		Privacy Policy	PP16	2877
Policy Guide 10	PG10	Terms and Conditions	PP17	2681
Policy Guide 11	PG11	Privacy Policy	PP18	394
		Disclaimer	PP19	1048
		Safety Guidelines	PP20	661
		Terms and Conditions	PP21	6520

Policy Guide 12	PG12	Privacy Policy	PP22	616
Policy Guide 13	PG13	Disclaimer	PP23	222
Policy Guide 14	PG14	Terms and Conditions	PP24	1982
		Privacy Policy	PP25	377
Policy Guide 15	PG15	Privacy Policy	PP26	535
		Terms and Conditions	PP27	1165
Policy Guide 16	PG16	Disclaimer	PP28	356
Policy Guide 17	PG17	Privacy Policy	PP29	1954
		Terms and Conditions	PP30	2183
Policy Guide 18	PG18	Terms and Conditions	PP31	469
Policy Guide 19	PG19	Terms and Conditions	PP32	3375
		Privacy Policy	PP33	738
Policy Guide 20	PG20	Terms and Conditions	PP34	5031
		Privacy Policy	PP35	453
Policy Guide 21	PG21	Terms and Conditions	PP36	611
		Privacy Policy	PP37	992
Policy Guide 22	PG22	Privacy Policy	PP38	1255
Policy Guide 23	PG23	Terms and Conditions	PP39	3094
		Privacy Policy	PP40	973
Policy Guide 24	PG24	Privacy Policy	PP41	525
		Terms and Conditions	PP42	2759
Policy Guide 25	PG25	Privacy Policy	PP43	948
		Disclaimer	PP44	1396
		Terms and Conditions	PP45	2104
Policy Guide 26	PG26	Terms and Conditions	PP46	3722
		Privacy Policy	PP47	388
Policy Guide 27	PG27	Terms and Conditions	PP48	545
Policy Guide 28	PG28	Privacy Policy	PP49	2119
		Safety Guidelines	PP50	819
		Terms and Conditions	PP51	2352
Policy Guide 29	PG29	Terms and Conditions	PP52	2312
		Privacy Policy	PP53	594
Policy Guide 30	PG30	Terms and Conditions	PP54	821
		Disclaimer	PP55	314

Table III: Readability Indexes for Different Web-Pages of Policy Guides of Matrimonial Sites

APG#	APP#	Character-based			Syllable-based			
		ARI	CLI	Average	FKGL	GFI	SMOGI	Average
PG1	PP1	7.7	13.5	10.6	7.9	10.4	8.5	8.93
PG2	PP2	9.4	13.1	11.25	10.3	12.1	9.3	10.57
	PP3	11.8	16	13.9	12.7	12	10.5	11.73
PG3	PP4	10	13	11.5	10.9	13.1	10.1	11.37
	PP5	10.1	13.3	11.7	11.2	13.3	10.3	11.6
PG4	PP6	10	15.8	12.9	11.2	10.7	9.5	10.47
PG5	PP7	7.6	11.2	9.4	8.6	10.9	8.2	9.23
PG6	PP8	13.4	12.9	13.15	13.6	16.4	12.3	14.1
	PP9	12.7	14.2	13.45	12.9	15.8	12.1	13.6
	PP10	8.4	11.8	10.1	9.1	11.4	8.5	9.67
	PP11	13	14.6	13.8	13.2	15.4	12.3	13.63
PG7	PP12	10.4	20.1	15.25	13.6	12.6	7.6	11.27
	PP13	9.8	13.1	11.45	10.6	12.3	9.5	10.8
PG8	PP14	11.8	13.4	12.6	12.2	14	11.1	12.43
PG9	PP15	9.2	13.3	11.25	10.2	12.6	10	10.93
	PP16	9.2	13.3	11.25	10.2	12.6	10	10.93
PG10	PP17	9.2	13.5	11.35	10.4	11.9	9.4	10.57

PG11	PP18	6.8	11.8	9.3	7.9	10	8	8.63
	PP19	12.5	15.4	13.95	13.4	16.4	12.6	14.13
	PP20	9.8	13.4	11.6	10.3	13	9.9	11.07
	PP21	12.7	13.4	13.05	13.1	15.1	11.5	13.23
PG12	PP22	8.5	11.8	10.15	9.4	11.9	9	10.1
PG13	PP23	10.7	14.5	12.6	11.2	13	10.5	11.57
PG14	PP24	9.7	15.3	12.5	10.3	10.6	9.3	10.07
	PP25	10	12.8	11.4	10.3	12.3	9.6	10.73
PG15	PP26	9.5	14.9	12.2	10.8	10.4	8.5	9.9
	PP27	11.2	15.7	13.45	12.5	13.5	10.7	12.23
PG16	PP28	7.3	12.7	10	8.7	9.6	8.8	9.03
PG17	PP29	10.9	14	12.45	10.8	12.8	9.8	11.13
	PP30	13.3	14.4	13.85	13	15	11.7	13.23
PG18	PP31	3.7	10.8	7.25	6.5	7.8	6.4	6.9
PG19	PP32	9.8	13.5	11.65	10.7	13.3	10	11.33
	PP33	8.4	14.6	11.5	9.3	11.5	8.5	9.77
PG20	PP34	10.4	20.1	15.25	13.6	12.6	7.6	11.27
	PP35	9.8	13.1	11.45	10.6	12.3	9.5	10.8
PG21	PP36	12.8	13.4	13.1	13	15.5	11.8	13.43
	PP37	9.3	14	11.65	9.8	11.3	9	10.03
PG22	PP38	11.6	14.7	13.15	11.4	13.7	10.4	11.83
PG23	PP39	9.5	14.2	11.85	10.7	12.5	9.9	11.03
	PP40	10.3	14.3	12.3	10.4	12.2	10.1	10.9
PG24	PP41	10.1	13	11.55	11	13	10	11.33
	PP42	10	14.7	12.35	10.9	11.9	9.9	10.9
PG25	PP43	12.7	13.6	13.15	13.1	16	12.1	13.73
	PP44	8.8	13.6	11.2	10.5	12.1	10.1	10.9
	PP45	14.8	12.1	13.45	14.5	16.8	12.4	14.57
PG26	PP46	12.1	13.1	12.6	12.7	15.1	11.7	13.17
	PP47	10	13.3	11.65	10.8	13.4	10.3	11.5
PG27	PP48	10.7	11.2	10.95	10.8	13	9.7	11.17
PG28	PP49	8.3	14.7	11.5	10	11.8	8.6	10.13
	PP50	8.1	12.6	10.35	8.5	10.9	8.2	9.2
	PP51	9.9	14.9	12.4	11.5	12.3	9.9	11.23
PG29	PP52	10.7	13.4	12.05	11.2	14	10.8	12
	PP53	9.4	13.5	11.45	9.8	11.7	9.3	10.27
PG30	PP54	9.7	13.4	11.55	10.3	11.6	9.2	10.37
	PP55	9	14.4	11.7	10.1	11.1	8.7	9.97

IV. Results and Findings

Table III presents the values of five readability indices for 55 webpages of 30 matrimonial websites, in a tabular format. Table III also presents the average of character-based indices as well as the average of syllable-based readability indices. Both the average values were rounded to two places of decimal for simplicity of calculations. Owing to minor rounding off for the floating point values, minor dilutions were introduced in the data. But these dilutions were of statistically irrelevant magnitude for the present work.

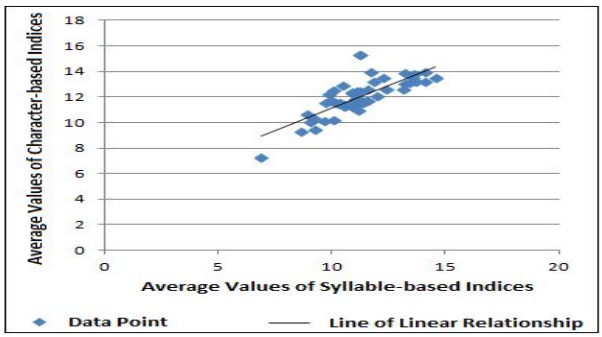

Figure 1: Mapping of Vectors Containing Averages of Character and Syllable-based Indices

Based on the data presented in Table III, the average values, rounded to two decimal places, for average of character-based indexes and average of syllable-based indices were found to be 11.97 and 11.18, respectively. This was the first result that proved that there is no difference of grade levels provided in the form of readability indexes by readability tests which are character-based and syllable-based. As the length of web-pages of different policy guides of different web-sites varies in terms of count of characters, syllables, words, etc., here interpretation of average values only is presented. Further, these average values being based on multiple inputs, are more generalized in nature.

In order to have a final confirmation of these results, the vector of average values for character-based indices on Y-axis was plotted against the average values for syllable-based indices on X-axis. The results were close clustering of points around the trend line, indicating that there is rarely any difference between the character-based and syllable-based indices. Even the difference occurring may be actually due to the role of outliers. This data is graphically presented in Figure 1.

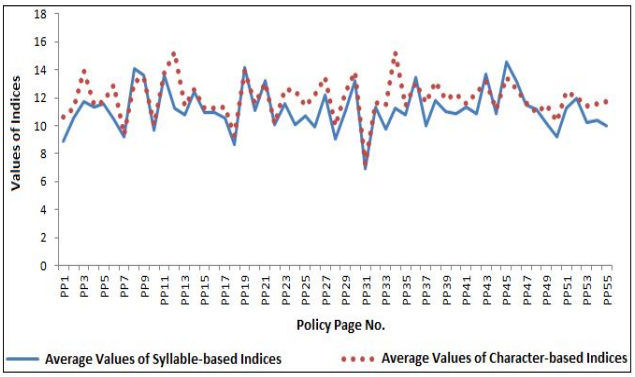

Figure 2: Plotting of Vectors Containing Averages of Character and Syllable-based Indices Against Policy Webpages

The hypothesis that there is no major difference between the readability score metric values provided by character-based and syllable-based readability tests was put to test by finding the Pearson Product-moment correlation coefficient. The value of this coefficient between the arrays of averages of indices of both bases was found to be 0.763657288, rounded off to 0.76, which is statistically an acceptable value of positive correlation between the value sets of two variables. Hence, the hypothesis that there is no difference between readability scores of set of character-based and set of syllable-based indices was accepted. Further, the vectors containing averages of character-based and syllable-based indices were also plotted together against the webpages. The close plotting of the graphical lines for both vectors on the chart also indicated their analogical values. This data is graphically presented in Figure 2. This was also to support the hypothesis that there is no major difference between the set of character-based and set of syllable-based sets of readability indices for estimation of comprehension ease of text documents.

However, the comparison of a single index of character-based set with each index of syllable-based set yielded that only ARI correlated well with FKGL (0.93), GFI (0.91) and SMOGI (0.88). The results of CLI for FKGL, GFI and SMOGI respectively with values of 0.497, 0.096 and -0.042 were discouraging. After these inter-set comparisons, the intra-set comparison of ARI with CLI yielded Pearson Product-moment Correlation Coefficient value of 0.27. Similarly, the coefficient values for {FKGL, GFI}, {FKGL, SMOGI} and {GFI, SMOGI} pairs were 0.85, 0.74 and 0.91, respectively.

Conclusion

The research works in past have proved that readability does play a role concerning comprehension of text under study. This paper presented a study, design, implementation of experiment and analysis of the results obtained for comparison of average of two character-based indices namely ARI and CLI with the average of three syllable-based indices namely FKGL, GFI and SMOG. Readability formulas can assist in estimating the reading level of materials by offering an objective measure of text difficulty. Factors such as color, font and other formatting characteristics of the web-pages were not the focus of current work. The 55 webpages of 30 matrimonial websites were used for the experiment and it is concluded that the average of character-based indexes and average of syllable-based indices is 11.97 and 11.18, respectively. There is a strong clustering of points around the trend line and the Pearson Product-moment correlation coefficient for values of averages of character-based and syllable-based indices is 0.76. Finally, the graphical lines for vectors of averages of character-based and syllable-based indices, on the chart also indicates that there is no major difference between the character-based and syllable-based sets of readability indices for estimation of comprehension ease of text documents. For the case of the individual indices of both the sets, only ARI has been found to have good inter-set correlation with the three indices considered for syllable-based tests. The intra-set comparison of syllable-based indices was much better than the intra-set comparison of character-based indices.

Even though the current results are best reported on the domain of web-pages of matrimonial websites under consideration, the author argues that the same results also hold true for another set of matrimonial sites as well as for any set of websites. The present work was solely an academic research work. This work neither defames nor promotes any matrimonial website. The work is also not intended to advertise matrimonial

websites as being the key source of marriages. This research work is not intended for any commercial value but advocates that there is a direct positive and linear correlation between the grade levels provided by character-based and syllable-based readability tests for policy guides of matrimonial websites and hence they could be used interchangeably.

References

[1] Dating Sites Reviews, "Internet Population Passes One Billion", Available: http://www.datingsitesreviews .com/article.php?story=Internet-Population-Passes-One-Billion

[2] Incisive Interactive Marketing LLC, "Stats - Web Worldwide", July 27, 2006. Available: http://www.clickz.com/stats/ web_worldwide

[3] Gyongyi Z. and Garcia-Molina H., "Web Spam Taxonomy", *Proceedings of First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb, 2005)*, April 2005, Chiba, Japan

[4] RFP Evaluation Centers, "What are Readability Statistics?", Available: rfptemplates.technologyevaluation.com/What-are-Readability-Statistics.html

[5] editcentral.com, "Introduction: On Writing – Style and Diction", Available: http://www.editcentral.com/gwt1/EditCentral.html

[6] Wikipedia, the free encyclopedia, "Pearson Product-moment Correlation Coefficient", Wikimedia Foundation Inc. Available: http://en.wikipedia.org/wiki/ Pearson_product-moment_correlation_coefficient

[7] Meade C. D., Byrd J.C. and Lee M., "Improving Patient Comprehension of Literature on Smoking" in *Proceedings of American Journal of Public Health (AJPH)*, ISSN: 0090-0036, 79(10), 1989, pp. 1411-1412

[8] McNamara D.S., Louwerse M.M., and Graesser A.C., "Coh-Metrix: Automated Cohesion and Coherence Scores to Predict Text Readability and Facilitate Comprehension", University of Memphis, Available: csep.psyc.memphis.edu/pdf/IESproposal.pdf

[9] Saini J. R. and Desai A. A., "Self Learning Taxonomical Classification System Using Vector Space Document Analysis Model For Web Text Mining In UBE", *Ph. D. Thesis.*, Department of Computer Science. VNSGU, Surat, 2009

[10] Saini J. R. and Desai A. A., "A Textual Analysis of Digits Used for Designing Yahoo-group Identifiers", published in *The IUP Journal of Information Technology*, ISSN: 0973-2896, 6(2), 2010, pp. 34-42

[11] Saini J. R. and Desai A. A., "Structural Analysis of Username Segment in Email Addresses of MCA Institutes of Gujarat State", published in *The IUP Journal of Information Technology*, ISSN: 0973-2896, 6(3), 2010, pp. 43-50

[12] Saini J. R. and Desai A. A., "A Classification of Character Usage in Unique Addresses Employed for Accessing Yahoo! Groups Service", published in *Karpagam Journal of Computer Science*, ISSN: 0973-2926, 12(1), 2011, pp. 233-240

[13] Belloni L.F. and Jongsma E.A., "The Effects of Interest on Reading Comprehension of Low-Achieving Students", in *Proceedings of Journal of Reading*, 22(2), 1978, pp. 106-109

[14] Saini J. R., "Analyzing Comprehension Ease of Policy Guides of Dating Sites Using Readability Statistics", published in *Journal of SCI-TECH Research*, ISSN: 0974-9780, 2(2), 2011, pp. 05-12

[15] Saini J. R., "Web Text Mining Through Readability Metrics for Evaluation of Understandability of Policy Guides of Matrimonial Websites", 2014, submitted for publication

[16] Saini J. R., "Estimation of Comprehension Ease of Policy Guides of Matrimonial Websites Using Gunning-Fog, Coleman-Liau and Automated Readability Indices", 2014, submitted for publication

[17] Mateus K., "Five Tips for Improving Website Readability and Content Density", Available: http://www.mequoda.com/articles/website-design/five-tips-for-improving-website-readability-and-content-density

[18] Sedlak W., "The Importance of Readability in Good Website Design", Available: http://ezinearticles.com/?The-Importance-of-Readability-in-Good-Website-Design&id=2591054

[19] Wikipedia, the free encyclopedia, "Matrimonial Website", Wikimedia Foundation Inc. Available: http://en.wikipedia.org/wiki/Matrimonial_website

Dr. Jatinderkumar R. Saini is Ph.D. from VNSGU, Surat. He secured First Rank in all three years of MCA and has been awarded Gold Medals for this. He is IBM Certified Database Associate (DB2) as well as IBM Certified Associate Developer (RAD). Associated with nearly 45 countries, he has been the Member of Program Committee for more than 50 International Conferences (including those by IEEE and Springer) and Editorial Board Member or Reviewer for more than 30 International Journals (including many those with Thomson Reuters Impact Factor). He has more than 55 research paper publications, many indexed by Elsevier, and nearly 20 presentations in reputed International and National Conferences and Journals. He has achieved more than 30 International Certifications from various universities of countries like U.S.A. and U.K. He has been listed in many distinguished Biographical Directories by many countries. He is member of ISTE, IETE, ISG and CSI. He is recognized PhD Supervisor and a member of Board of Studies of various universities across India. Currently he is working as Associate Professor and Director I/C at Narmada College of Computer Application, Bharuch, Gujarat, India. He is also Director (Information Technology) at Gujarat Technological University, Ahmedabad (GTU)'s A-B Innovation Sankul.