DEVELOPMENT OF A READABILITY INDEX ATTUNED TO THE NEW ENGLISH COURSE OF STUDY IN JAPAN

Toshiaki Ozasa[1], George R. S. Weir[2] and Masayasu Fukui[1]

[1]Faculty of Business Administration, Fukuyama Heisei University, Fukuyama, Japan

[2]Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK

tozasa@heisei-u.ac.jp, george.weir@cis.strath.ac.uk, fukui@heisei-u.ac.jp

Abstract

For several years, we have been engaged in evaluating techniques for estimating readability. Arising from this work, there have been several iterations of readability indices attuned to the English textbooks employed in Japanese high schools for the English Course of Study. In this paper, we briefly review approaches to readability measurement before detailing our project that aims to develop new readability indices that are attuned to the new English teaching system of Japan. Our earlier approach to the Japanese English teaching context employed regression analysis to develop a linear measure in which sentence length, word length, word difficulty and idiom difficulty are independent variables and the textbook year of passages is a dependent variable. This linear function (Diff) proved to be broadly effective as a guide to reading level for English texts in the Japanese school context and was released for use as 'the Ozasa-Fukui Year Level'. From 2012, changes to the English Course of Study in Japan made it necessary to update this English readability index. To this end, our earlier approach was enhanced to provide a non-linear formula in which an experts' readability judgment of English sentences is employed as a dependent variable and combined with the previously developed linear function (Diff) as an independent variable. The resultant computation yielded a non-linear function expression: NewDiff = a/(1+b*exp(-c*Diff))+0.9 in which we have the following component values: a=3.8593, b=766.9372, c=2.5709, Diff = 0.0863*Words/S + 0.2943*Syllables/W + 0.6332*WordDiff/W + 0.0665*IdiomDiff/S + 0.5366. Applying this new measure, we determined that the coefficient of determination for this new formula (r^2) was 0.8236. Thereby, indicating a high degree of fit to the target data.

Keywords: Measuring Readability, English Course of Study of Japan, Least squares method, Non-linear analysis, Ozasa-Fukui Year Level

1 INTRODUCTION

Since 2007, we have been engaged in developing more sophisticated approaches to readability [1,2,3]. In the Japanese context, we have focussed on indices for measuring readability based on the textual data of the English textbooks for junior and senior high schools of Japan [4-12]. In an attempt to develop indices for measuring the readability of EFL texts used in the Japanese formal education context, we developed three readability indices, Ozasa-Fukui Year Level, Ver. 1, 2, and 2.1, which were attuned to the English Course of Study of Japan (effective 2002-2011 for junior high school, 2003-2012 for senior high school). The main features of the three readability indices developed under this scheme are outlined in Table 1.

In developing Ver. 1, a linear readability index was developed through a series of multiple regression analyses using sentence length, word length, textbook-based word difficulty and textbook-based idiom difficulty as independent variables and year level of EFL textbook sentences as a dependent variable. The analyses yielded a linear function expression with a low prediction rate (r^2= 0.387) for the index of Ver. 1.

In developing Ver. 1, the criterion readability measure used as a dependent variable for the linear analyses was year level of the textbooks (five levels), in which a sentence which appears in a textbook was automatically allotted a year level of that particular textbook. A detailed critical examination of the results of the multiple regression analyses, however, revealed that the criterion was not accurate enough as a dependent variable to guarantee a high prediction rate for the index.

In developing Ver. 2, based on this criticism, an attempt was made to develop a new readability criterion as a dependent variable for the second-step, nonlinear analysis. First, in the first-step analysis, a multiple regression analysis, a linear analysis, was computed, using sentence length, word length, word difficulty and idiom difficulty as independent variables and the year (mean values) of all the passages of all of the English textbooks used in the analysis as a dependent variable, which yielded a linear expression (Diff). Then, in the second-step analysis, non-linear analyses were computed, using the new, experts' readability judgment as a dependent variable and the linear function (Diff) as an independent variable. The measure used for the dependent variable was a new criterion specifically developed for this analysis, which is based on the intuitive judgment of 126 sentences selected from two sets of the Japanese EFL textbooks (10 volumes in total) by three experienced Japanese EFL teachers who are well familiar with the EFL textbooks and teaching at middle-grade schools in Japan. The analyses yielded a linear function expression with the prediction rate (r^2) of 0.824) for the index of Ver. 2, which proved to be a drastic improvement in the quality of the index.

Table 1 Versions of O-F Year Level Compared

Version	r^2	Dependent variable			Independent variable	Analysis
		Evaluation	Level Number	Datum size		
1	0.387	Objective	5	Big[*1]	4[*2]	linear
2	0.824	Empirical (3)	50	126	4	nonlinear
2.1	0.822	Empirical (3)	50	916	4	nonlinear
3.1nh	0.7902	Empirical (3)	40	546	4	nonlinear
3.1nc	0.8373	Empirical (3)	40	151	4	nonlinear
3.1nhnc	0.7502	Empirical (3)	40	697	4	nonlinear

[*1] All sentences contained in 15 vols. of the textbooks.

[*2] Sentence length + word length + textbook word difficulty + textbook idiom difficulty

In developing Ver. 2.1, the technique and procedure employed in the analysis were basically the same as those used for Ver. 2, only difference being in the size and kinds of the dependent variable used in the second stage of the analysis; the number of the data used in the second analysis was 916 and the kinds of English textbooks used were three sets of representative Japanese EFL textbooks (15 volumes in total). The analysis yielded a non-liner function expression (NewDiff) which proved to be nearly as powerful in explanatory power (r^2=0.822) as the previously obtained nonlinear function of Ver. 2.

In 2012 a new educational system was effected; the newly revised English Course of Study was put into practice in 2012 for junior high school and 2013 for senior high school. To cope with this situation, we launched a new project aiming to develop new readability indices that are exactly attuned to the newly revised English Course of Study. The basic analysis procedure was the same as that of Ver. 2 and 2.1. In the first step, a multiple regression analysis, a linear analysis, was computed, using sentence length, word length, word difficulty and idiom difficulty as independent variables and the textbook year of passages as a dependent variable, which yielded a linear function (Diff). In the second step, non-linear analyses were computed, using the empirical judgment of English sentences as a dependent variable and the linear function (Diff) as an independent variable.

The analyses, reported in [11] and [12], yielded three non-linear function expressions, Ver. 3.1nh, Ver. 3.1nc and Ver. 3.1nhnc (See Table 1). In developing Ver. 3.1nh, experts' readability evaluation of 546 English sentences selected from New Horizon English Course, 1, 2 & 3, and Prominence English Communication, 1 was used as a measure for the dependent variable, which yielded a nonlinear function expression with the prediction rate (r^2) of 0.7902. In developing Ver. 3.1nc, experts' readability evaluation of 151 English sentences selected from New Crown English Series, 1, 2 & 3 and Crown English Communication, 1 was used as a measure for the dependent variable, which yielded a nonlinear function expression with the prediction rate (r^2) of 0.8373. In developing Ver. 3.1nhnc, experts' readability evaluation of 697 English sentences selected from New Horizon English Course, 1, 2 & 3, Prominence English Communication, 1, New Crown English Series, 1, 2 & 3, and Crown English Communication, 1 was used as a measure for the dependent variable, which yielded a nonlinear function expression with the prediction rate (r^2) of 0.7502.

The prediction or explanation rates of the functions yielded, i.e., 0.7902 for Ver. 3.1nh, 0.8373 for Ver. 3.1nc, and 0.7502 for Ver. 3.1nhnc. It was concluded that these prediction rates were not as high as those of the former versions, i.e., 0.824 for Ver. 2 and 0.822 for Ver. 2.1.

2 AIM

The present paper aims to improve the quality of the recently developed readability indices in prediction rates, i.e., Ver. 3.1nh, Ver. 3.1nc and Ver. 3, 1nhnc, developing new readability indices of higher quality in prediction or explanation that are attuned to the newly revised English Course of Study that were effected in 2012 and 2013. This is the third interim report of the new project that purports to develop new readability indices for the new course of study.

2.1 Method

In the analysis, as the first step, a multiple regression analysis, a linear analysis, was computed, using sentence length, word length, word difficulty and idiom difficulty as independent variables and the year (mean values) of all the passages of all of the English textbooks used in the analysis as a dependent variable, which yielded a linear expression (Diff).

As the second step, non-linear analyses were computed, using the experts' judgment of English sentences as a dependent variable and the linear function (Diff) as an independent variable. The measure used for the dependent variable was a new criterion specifically developed for this analysis, which is based on the intuitive judgment of sentences selected from the two sets of the Japanese EFL texts by three experienced Japanese EFL teachers who are well familiar with the English textbooks and teaching at middle-grade schools in Japan. The textbooks used were as follows.

New Crown English Series, 1, 2 & 3. [Junior high school] 2012, Tokyo:Sanseido.
Crown English Communication, *1*. [Senior high school] 2013. Tokyo: Sanseido.
New Horizon English Course, 1, 2 & 3, [Junior high school] 2012, Tokyo: Tokyo Shoseki.
Prominence English Communication, 1. [Senior high school] 2013. Tokyo: Tokyo Shoseki.

Several linear and nonlinear analyses were computed, using the newly developed subjective criterion measure as a dependent variable and the previously obtained linear function (Exp. 1) as an independent variable. The computation was carried out using Fukui's College Analysis, a statistics computer program developed by Masayasu Fukui for social studies.

3 RESULTS AND DISCUSSION

3.1 Development of Ver. 3.2nh

The procedure employed in developing Ozasa-Fukui Year Level, Ver. 3.2nh, was the same as that of Ver. 3.1nh and Ver. 3.1nhnc. First, a multiple regression analysis, a linear analysis, was computed, using sentence length, word length, word difficulty and idiom difficulty as independent variables and the mean values of the school years of all the passages of the English textbooks as a dependent variable. The textbooks used for the analysis were New Horizon English Course, 1, 2 & 3 and Prominence English Communication, 1. The following is the results of the regression analysis.

Object Variable		Year
Explanatory Variables		Words/S, Syllables/W, WordDiff/W, IdiomDiff/S
Number of Data		915
Regression Exp.		
Diff = 0.0710*Words/S+0.3640*Syllables/W+1.1846*WordDiff/W+0.0637*IdiomDiff/S+0.0364		
Coeff. Det.	R^2	0.4261
Multi. Corr.	R	0.6528
Adjusted Corr.	R'	0.6508

As is clear in the analysis data, the computation yielded the following regression expression (Diff).

Diff = 0.0710*Words/S+0.3640*Syllables/W+1.1846*WordDiff/W+0.0637*IdiomDiff/S+0.0364 (Diff)

As a next step, non-linear, least squares analyses were computed, using the experts' readability judgment of 157 English sentences as a dependent variable and the resultant linear function (Diff) as an independent variable. This computation yielded two non-linear function expressions, the Gompertz solution and the Logistic solution. The Gompertz solution is shown in the following analysis result.

Object Variable	Year
Expression	NewDiff = a*exp(-b*c^Diff)+1

Optimal Solution		
A		3.8918
B		48.2661
C		0.2191
Num. of Cases		157
Num. of Solutions		1
Least Sq. Value		25.61430
Obs/Pred	R	0.9563
Obs/Pred	R^2	0.9144

As is clear from this analysis data, the computation yielded the following nonlinear function expression (NewDiff).

NewDiff = 3.8918*exp(-48.2661*0.2191^Diff)+1
(Diff=0.0710*Words/S+0.3640*Syllables/W+1.1846*WordDiff/W+0.0637*IdiomDiff/S+0.036)

The explanatory rate of this solution (r^2) of 0.9144, is much higher than that of Ver. 3.1nh (0.7902). Presumably, this high rate was influenced by the veteran teachers' detailed, close re-examination and improvement of the former readability evaluation measure used for the development of Ver. 3.1nh. Fig. 1 shows the dispersion of the Gompertz predictions and the line of NewDiff (Gompertz prediction).

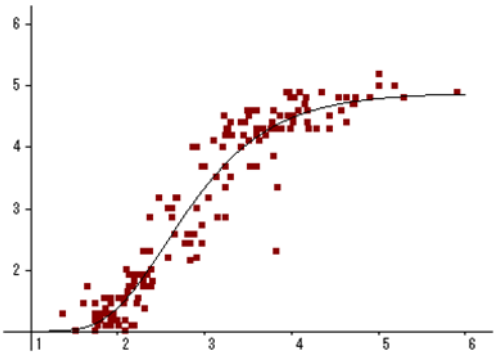

Figure 1: Prediction of Gompertz solution

The second solution, the Logistic solution is shown in the following analysis result.

Object Variable		Year
Expression		NewDiff = a/(1+b*exp(-c*Diff))+0.9
Optimal Solution		
A		3.8167
B		616.9787
C		2.3231
Num. of Cases		157
Num. of Solutions		1
Least Sq. Value		25.9428
Obs/Pred	R	0.9556
Obs/Pred	R^2	0.9133

As is clear from this analysis data, the computation yielded the following nonlinear prediction expression (NewDiff).

NewDiff =3.8167 /(1+616.9787*exp(-2.3231*Diff))+0.9
(Diff=0.0710*Words/S+0.3640*Syllables/W+1.1846*WordDiff/W+0.0637*IdiomDiff/S+0.036)

The explanatory rate of this solution (r^2) proved to be 0.9133, which is also exceptionally high, much higher than that of Ver. 3.1nh (0.7902). Presumably, this difference was achieved by the veteran teachers' detailed, close re-examination and improvement of the former readability evaluation criterion used for the development of Ver. 3.1nh. Fig. 2 shows the dispersion of the predictions and the line of NewDiff (Logistic prediction).

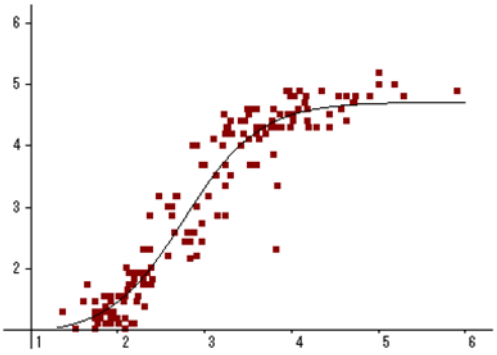

Figure 2: Prediction of Logistic solution

Since the Gompertz solution proved to be slightly more efficient in explanation than the Logistic solution, the Gompertz expression was adopted as the index function of the Ozasa-Fukui Year Level, Ver. 3.2nh.

3.1 Development of Ver. 3.2nhnc

Since the results of the least squares analysis for the development of the Ver. 3.2nh were satisfactory to our expectation, an attempt was made to integrate the data of Ver. 3.2nh and those of Ver. 3.1nc and to develop an readability index of higher validity, Ver. 3.2nhnc. The basic process of the analysis was the same as that used for the development of Ver. 3.2nh.

First, a multiple regression analysis, a linear analysis, was computed, using sentence length, word length, word difficulty and idiom difficulty as independent variables and the mean values of the school years of all the passages of the English textbooks as a dependent variable. The textbooks used for the analysis were New horizon English course, 1, 2 & 3, Prominence English Communication, 1 , New crown English series, 1, 2 & 3 and Crown English Communication, 1. The following gives the results of the regression analysis.

Object Variable		Year
Explanatory Variables		Words/S, Syllables/W, WordDiff/W, IdiomDiff/S
Number of Data		2728
Regression Exp.		
Diff = 0.0863*Words/S+0.2943*Syllables/W+0.6332*WordDiff/W+0.0665*IdiomDiff/S+0.5366		
Coeff. Det.	R^2	0.3386
Multi. Corr.	R	0.5819
Adjusted Corr.	R'	0.5811

From this analysis data, the computation yielded the following regression expression (Diff).

$$\text{Diff} = 0.0863*\text{Words/S}+0.2943*\text{Syllables/W}+0.6332*\text{WordDiff/W}+0.0665*\text{IdiomDiff/S}+0.5366 \quad \text{(Diff)}$$

For the next phase, non-linear, least squares analyses were computed, using the experts' judgment of 308 English sentences as a dependent variable and the resultant linear function (Diff) as an independent variable. The computation yielded two non-linear function expressions, the Gompertz solution and the Logistic solution. The Gompertz solution is shown in the following analysis result.

Object Variable		Year
Expression		NewDiff = a*exp(-b*c^Diff)+1
Optimal Solution		
A		3.9597
B		52.7712
C		0.1921
Num. of Cases		308
Num. of Solutions		1
Least Sq. Value		96.9251
Obs/Pred	R	0.9060
Obs/Pred	R^2	0.8208

As is clear from this analysis data, the computation yielded the following nonlinear prediction expression (NewDiff).

NewDiff = 3.9597*exp(-52.7712*0.1921^Diff)+1
(Diff=0.0863*Words/S+0.2943*Syllables/W+0.6332*WordDiff/W+0.0665*IdiomDiff/S+0.536)

The explanatory rate of this solution (r^2) was 0.8208, which is not as high as that of Ver. 3.2nh, but which is much higher than that of Ver. 3.1nhnc (0.7502). Presumably, this difference was achieved by the improvement of the former readability evaluation criterion used for the development of Ver. 3.1nhnc. Fig. 3 shows the dispersion of the Gompertz predictions and the line of NewDiff (Gompertz prediction).

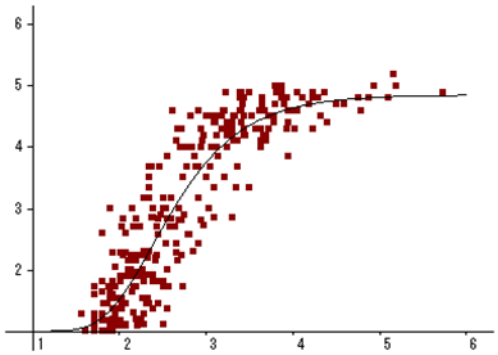

Figure 3: Prediction of Gompertz solution

The second least squares solution, the Logistic solution, is shown in the following analysis result.

Object Variable		Year
Expression		NewDiff = a/(1+b*exp(-c*Diff))+0.9
Optimal Solution		
A		3.8593
B		766.9372
C		2.5709
Num. of Cases		308
Num. of Solutions		1
Least Sq. Value		95.3330
Obs/Pred	R	0.9075
Obs/Pred	R^2	0.8236

Based upon this analysis data, the computation yielded the following nonlinear prediction expression (NewDiff).

NewDiff = 3.8593/(1+766.9372*exp(-2.5709*Diff))+0.9
Diff=0.0863*Words/S+0.2943*Syllables/W+0.6332*WordDiff/W+0.0665*IdiomDiff/S+0.536)

The explanatory rate of this solution (r^2) proved to be 0.8236, which is not as high as that of Ver. 3.2nh, but which is much higher than that of Ver. 3.1nhnc (0.7502). Presumably, this difference was achieved by the improvement of the former readability evaluation criterion used for the development of Ver. 3.1nhnc. Fig. 4 shows the dispersion of the Logistic predictions and the line of NewDiff (Logistic prediction).

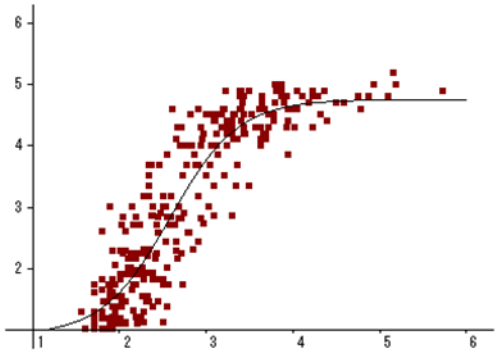

Figure 4: Prediction of Logistic solution

Since the function expression of the Logistic solution (0.8236) was higher than that of the Gompertz solution (0.8208) in explanation (r^2), the Logistic solution was formally adopted as the index of the Ver. 3.2nhnc at the present stage of this project.

4 CONCLUSION

In the present study, the following two function expressions were adopted as the indices of Ozasa-Fukui Year Level, Ver. 3.2nh and Ver. 3.2nhnc:

Ver. 3.2nh = 3.8918*exp(-48.2661*0.2191^Diff)+1
(Diff=0.0710*Words/S+0.3640*Syllables/W+1.1846*WordDiff/W+0.0637*IdiomDiff/S+0.036)
(r^2=0.9144)

Ver. 3.2nhnc = 3.8593/(1+766.9372*exp(-2.5709*Diff))+0.9
(Diff=0.0863*Words/S+0.2943*Syllables/W+0.6332*WordDiff/W+0.0665*IdiomDiff/S+0.5366)
(r^2=0.8236)

For practical purposes, the choice of a function expression depends on the purpose that a user has in mind when estimating the readability of English texts. When emphasis is placed on measuring readability in terms of New Horizon textbooks, Ver.3.2nh should be selected. On the other hand, when emphasis is placed on measuring the readability in general, Ver. 3.2nhnc should be selected.

REFERENCES

[1] Anagnostou, N. K. & Weir, G. R. S. (2006). From corpus-based collocation frequencies to readability measure. *ICT in the Analysis, Teaching and Learning of Languages* (ICTATLL Workshop Preprints), 33-46.

[2] Weir, G. R. S. & Ritchie, C. (2006). Estimating readability with the Strathclyde readability measure. ICT in the Analysis, Teaching and Learning of Languages (ICTATLL Workshop Preprints), 25-32.

[3] Anagnostou, N. K. & Weir, G. R. S. (2007). Average collocation frequency as an indicator of semantic complexity. *ICTATLL Workshop 2007 Preprints*, 43-48.

[4] Ozasa, T., Fukui, M., Hosokawa, M. & Miyoshi, F. (2007). Development of a readability index for Japanese learners of EFL (1): An interim report. *ICT in the Analysis, Teaching and Learning of Languages* (ICTATLL Workshop), 29-35.

[5] Ozasa, T, Weir, G. R. S. & Fukui, M. (2007). Measuring readability for Japanese learners of English. *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*, 122-125.

[6] Ozasa, T., Fukui, M., & Hosokawa, M. (2008). Development of a readability index attuned for Japanese EFL texts. *Proceedings of the 4th International ICT in the Analysis, Teaching and Learning of Languages (ICTATLL) Workshop*, 60-66.

[7] Ozasa, T., Weir, G. R. S. & Fukui, M. (2008). Toward a readability index for Japanese learners of EFL. *Conference Program of the 13th Conference of Pan-Pacific Association of Applied Linguistics*, 59-62.

[8] Fukui, M. & Ozasa, T. (2008). Methodology for developing an English readability index with word and idiom difficulty variables. (In Japanese), *Educational Information Studies*, 24 (3), 15-22.

[9] Ozasa, T., Abe, N., Hosaka, Y. & Fukui, M. (2009). Development of a user-friendly readability application software. *ICT in the Analysis, Teaching and Learning of Languages*, 41-49.

[10] Ozasa, T. & Weir, G. R. S. (2009). Evaluating a newly developed index for readability measurement of Japanese EFL textbooks. *Proceedings of the 14th Conference of Pan-Pacific Association of Applied Linguistics*, 95-100.

[11] Ozasa, T., Watanabe, K., Uenishi, K. & Hosaka, Y. (2013). Development of a new English readability measure based on 2013 Course Study: Development of Ozasa-Fukui Year Level, 3.1nh, (in Japanese), *PAAL Japan Journal, 1* (in print).

[12] Fukui, M., Ozasa, T. & Hosokawa, M. (2013). Development of new indices for measuring English readability based on the 2013 Course of Study of Japan: Developing Ozasa-Fukui Year Level, Ver. 3.1nhnc, (in Japanese), *Fukuyama Heisei University Business Administration Studies*, *10*.