Readability indices for the assessment of textbooks: a feasibility study in the context of EFL

Pascual Cantos Gómez
Universidad de Murcia, Spain
pcantos@um.es

Ángela Almela Sánchez-Lafuente
Universidad de Murcia, Spain
angelalm@um.es

Abstract

Readability indices have been widely used in order to measure textual difficulty. They can be useful for the automatic classification of texts, especially in language teaching. Among other applications, they allow for the previous determination of the difficulty level of texts without the need of reading them through. The aim of this research is twofold: first, to examine the degree of accuracy of the six most commonly used readability indices, and second, to present a new optimized measure. The main problem is that these readability indices may offer disparity, and this is precisely what has motivated our attempt to unite their potential. A discriminant analysis of all the variables under examination has enabled the creation of a much more precise model, improving the previous best results by 15%. Furthermore, errors and disparities in the difficulty level of the analyzed texts have been detected.

Keywords: Readability indices, text difficulty, EFL, EFL textbook, automatic classification of texts.

Resumen

Los índices de legibilidad se han utilizado de forma extensiva para determinar la dificultad textual, y pueden resultar muy útiles para la clasificación automática de textos, en especial en el ámbito de la enseñanza de lenguas. Entre otras de sus aplicaciones, está la de poder determinar la dificultad de texto sin necesidad de leerlo previamente. El objetivo de estudio es doble: por un lado, analizar el grado de precisión de los seis índices de legibilidad más utilizados, y por otro lado, partiendo de estos datos, intentar diseñar una nueva medida de legibilidad optimizada. El principal

problema es que estos índices pueden ofrecer disparidad, y es precisamente lo que ha motivado nuestro intento de unificar su potencial. Un análisis discriminante de todas las variables examinadas ha permitido la creación de un modelo mucho más preciso, mejorando los resultados previos en un 15%. Además de ello, es importante destacar que se han detectado errores y disparidades en el nivel de dificultad de los textos analizados.

Palabras clave: índices de legibilidad, dificultad textual, inglés como lengua extranjera, libro de texto de inglés, clasificación textual automática.

1. Introduction: formalizing text difficulty by virtue of readablility indices

Readability indices allow measuring how difficult it is to read a text based on its properties, by using constructs known to reflect complexity, such as average sentence length and number of complex words (Fry, 1968; Ash & Edgell, 1975). In the 1950s, these readability indices became increasingly popular, and researchers in the field devoted great effort to devising a substantial number of new formulae, since they can be useful for the automatic classification of texts, especially within language teaching.

Among other applications, readability indices allow for the previous determination of the difficulty level of texts without the need of reading them through. This is precisely what distinguishes readability formulae from comprehensibility tests, such as cloze tests: the former are determined only by the text itself, offering a value which indicates the complexity of the text only on the basis of quantitative elements, while the latter, first described by Taylor (1953), measures the comprehensibility of a text, that is to say, how understandable a text is to an actual reader. In other words, cloze tests give an actual measure of comprehension while readability formulae make a prediction. Precisely for this reason, even though they have been important in traditional readability research and readability formulae have been based on their results, comprehensibility tests have not been used in the present study, mainly quantitative in nature.

As this study is not intended to provide an extensive review of all the readability formulae, only a brief overview and description of the most commonly used readability indices is offered below.

1.1. Traditional Approaches to Readability

First of all, the Flesch/Flesch–Kincaid readability tests include two indices: the *Flesch Reading Easiness Score* and the *Flesch–Kincaid Grade Level*. The first system was devised by Rudolf Flesch in 1948. After several attempts at simplification (Farr, Jenkins, & Paterson, 1951; Kincaid, Fishburne, Rogers, & Chissom, 1975), this is the resulting formula:

$$FRE = 206.835 - 1.015 * \left(\frac{total\_words}{total\_sentences} \right) - 84.6 * \left(\frac{total\_syllables}{total\_words} \right)$$

In 1976, a revision commissioned by the U.S. Navy resulted in a modification of this index to generate a grade-level score, enabling the translation from the 0–100 score to a U.S. grade level. Nowadays, the ensuing formula is known as the *Flesch–Kincaid Grade Level*:

$$F\_KGL = 0.39 * \left(\frac{total\_words}{total\_sentences} \right) + 11.8 * \left(\frac{total\_syllables}{total\_words} \right) - 15.59$$

As can be seen, it uses the same core measures as the Reading Easiness test, namely word length and sentence length. However, the weighting factors are different, and the results of the two tests hence correlate inversely. In this way, a text with a relatively high score on the former test normally achieves a lower score on the latter.

A further index whose score corresponds to U.S. grade level is the *Gunning Fog Index*, or simply *Fog Index*. It was developed by R. Gunning (1952), becoming particularly popular owing to its easy calculation without a calculator (DuBay, 2004). *GFI* gets its index from mean sentence length (in words) and average number of complex words (words with three and more syllables):

$$GFI = 0.4 * \left(\frac{words}{sentence} \right) + 100 * \left(\frac{complex\_words}{words} \right)$$

Subsequently, *Automated Readability Index (ARI)* was worked out by Smith and Senter (1967) for the U.S. Army, and its validity on technical materials was proved by Smith and Kincaid (1970). The formula uses mean word length (in characters) and mean sentence length (in words):

$$ARI = 4.71 * \left(\frac{characters}{word} \right) + 0.5 * \left(\frac{words}{sentences} \right) - 21.43$$

In 1969, G. H. McLaughlin published *SMOG (Simple Measure of Gobbledygook)* in an attempt to make *Gunning Fog Index* calculation even easier. Indeed, in his work the author describes it as "laughably simple" (McLaughlin, 1969, p. 639). It is based upon the conviction that word length and sentence length are to be multiplied rather than added. The formula used at present is the following one:

$$SMOG\_grade = 1.043\sqrt{30 \times \frac{number\_of\_polysyllables}{number\_of\_sentences}} + 3.1291$$

where polysyllable count refers to the number of words of more than two syllables. The resulting score corresponds to the years of education needed to thoroughly understand a given piece of writing.

Finally, the *Coleman-Liau Index* was devised by Coleman and Liau (1975). Like the *ARI*, this measure relies on characters instead of syllables per word, which, as commented on above, is not the trend in readability indices. A further point of similarity between the *ARI* and the *CLI* which is also shared by the Flesch–Kincaid readability tests and the *GFI* is that the ensuing score stands for U.S. grade level. The *CLI* is calculated with the following formula:

$$C\_LI = 5.89 * \left(\frac{characters}{words}\right) + 29.5 * \left(\frac{sentences}{words}\right) - 15.8$$

1.2. Current Research in Readability

One of the main criticisms of the features used for the calculation of traditional readability indices is that they are considered to be linguistically shallow. However, as DuBay (2004) puts it, they are surprisingly effective and widely used at the present moment. Some attempts to combine classical features with other linguistic components for the prediction of text complexity have been recently made. Such is the case of Vajjala and Meurers (2012; 2013; 2014a; 2014b), Crossley, Yang, and McNamara (2014), Flor and Beigman (2014), and Fitzgerald et al. (2015), among others, who take into account language-specific morphological features or the quantification of coherence and cohesion in a text.

Some researchers have tried to validate traditional readability indices for EFL use, like Brown (1998) and Greenfield (1999). The former examined their performance administering cloze tests to 2,300 Japanese learners of EFL and comparing the results with scores predicted by traditional readability indices. Greenfield measured the performance of 200 Japanese college students on cloze tests, this time on a selection of academic passages. Interestingly enough, these two studies yielded contradictory

results: while Greenfield (1999) found traditional formulae to be predictive of reading difficulty, this was not the case of Brown (1998). As Crossley, Allen, and McNamara (2011: 87) put it,

Greenfield (2004) argued that Brown's (1998) passage set was not sufficiently variable in difficulty and too difficult overall to provide a robust passage set for L2 learners. Overall, these studies offer some evidence that classic readability measures discriminate reading difficulty reasonably well for L2 students, but are limited to the appropriate academic texts for which they were designed and do not reach the level of accuracy achieved in L1 cross-validation studies (Greenfield, 1999).

Along these lines, Crossley et al. (2011) compared the classification potential of some of the traditional readability indices mentioned above to more modern readability formulae based on psycholinguistic and cognitive accounts of text processing in discriminating between levels of L2 reading texts, exploring which readability index best classifies text level. However, to our knowledge, no study has compared the performance of the whole set of traditional readability indices with the further purpose of optimizing the results.

No doubt, the level of usage of readability formulae in educational contexts has diminished hugely; yet they are still used heavily to judge the readability of medical patient education materials (e.g. Freda, 2005; Cronin, O'Hanlon, & O'Connor, 2011). However, the main critique of the use of these formulae is limited to the observation that there is no consensus as to which readability formula is best suited for assessing patient education materials. Guo, Zhang, and Zhai (2011) argued that it is preferable to use more than one readability method to improve the validity of the results. Thus, although they have their limitations, such as overemphasis on observable character/word counts, morphological, syllabic features, etc., they are becoming more popular than ever (see Guo, Zhang & Zhai, 2011: 103). It appears that, despite the critiques, readability formulae are still perceived to have a useful function in a number of fields. It was partly to re-examine this functionality that the present study was carried out.

The main trouble with using readability indices is their disparity, and this is precisely what has motivated this paper: attempting to unite their potential. It is certainly true that the limitations of these indices have provoked much discussion and debate, and that in the last decades of the 20th Century there was serious criticism on their extensive use in areas such as law, journalism or health care. Some representative instances of this scholarly controversy are Maxwell (1978) and Connaster (1999), who offered some reasonable alternatives to readability indices like usability testing. Nevertheless, as DuBay (2004:3) puts it, "although the alternatives are useful and even

necessary, they fail to do what the formulas do: provide an objective prediction of text difficulty".

2. Research goal

The aim of this investigation is twofold: first, to examine the accuracy of six of the most commonly used traditional readability indices: *Flesch Reading Ease*, *Flesch–Kincaid Grade Level*, *Gunning Fog, Automated Readability Index*, SMOG, and *Coleman-Liau*; and second, by means of the data obtained, to present a new optimized measure using *Discriminant Function Analysis*.

These six formulae have been chosen because they represent the traditional approach whose performance the present authors aimed to test on EFL materials. Readability indices like the Lorges and Dale-Chall formulae have been excluded from this study because they do not only use quantitative parameters such as average sentence length and number of different words, but also lists of the most common words in English, mainly subsets of the *Dale list of 3000* (Dale & Chall, 1948). Such parameters would entail an external element, and we were mainly interested in the combination of parameters which could be calculated from the text itself.

Although some comparative studies on readability indices (e.g. Crossley et al., 2011) suggest that the *Flesch–Kincaid Grade Level* index is a revision of the *Flesch Reading Ease* index in order to ease its interpretation, we have decided to keep both in our study as intermediate tests obtained spoke against this observation. Partial correlation depending on the linguistic level was only significant for B1 texts (0.99), but not for A2.1 (0.14), A2.2 (0.02) or B2 (-0.36); and this is also applicable to the overall correlation (0.14).

3. Methodology

3.1. Task and Procedures

In order to test the accuracy of the six readability indices mentioned above, the indices of 20 already graded texts have been calculated, five texts for each linguistic level from the coursebook series *Innovations* (Dellar & Walkley, 2005a; Dellar & Walkley, 2005b; Dellar, Walkley, & Hocking, 2004; Dellar, Hocking, & Walkley, 2004). *Innovations* is a five-level general English course for foreign students. For this research, we have only taken the first four books and randomly extracted five text samples for each linguistic level (A2.1, A2.2, B1 and B2).

The randomly chosen texts for each linguistic level have been arranged and analysed according to their chronological order in each of the textbooks they belong to. That is, the elementary/A2.1 text with OD-code=1 occurs in the textbook previous to the one with OD-code=2, etc. (see Table 1).

3.2. Data Analysis

Before calculating the six readability indices, we first obtained the essential quantitative counts for each text, necessary for the various readability indices calculations: number of characters, sentence count, number of complex words (word of more than two syllables), and syllable count. Except for the latter, all these parameters were calculated with *WordSmith Tools 6.0*. As for syllable count, a reliable piece of freeware has been used: *WordCalc*.

In addition, each text sample was typified with a reading-order difficulty code (*OD-Code*: 1 to 20), according to its occurrence sequence in the textbooks and its corresponding linguistic level code (LL-Code: 1 = elementary/A2.1, 2 = pre-intermediate/A2.2, 3 = intermediate/B1, and 4 = upper-intermediate/B2). The preliminary data of all 20 texts are given in Table 1 below. Intuitively, the order of the text samples in Table 1 corresponds to its sequence of appearance in the various textbooks. Therefore, we might assume that text with OD-code = 1 and LL-code = 1 is, in principle, easier to read than text with OD-code = 5 and LL-code = 1.

Table 1. Data summary

LL	OD-Code	LL-Code	Tokens	Characters	Sentences	Syllables	Complex words
A2.1	1	1	289	1161	37	312	5
A2.1	2	1	278	1112	25	330	7
A2.1	3	1	322	1426	29	399	12
A2.1	4	1	233	1014	41	270	4
A2.1	5	1	268	1104	21	320	3
A2.2	6	2	306	1174	24	347	12
A2.2	7	2	564	2089	43	608	6
A2.2	8	2	444	1772	27	482	8
A2.2	9	2	608	2453	44	676	15
A2.2	10	2	661	3062	44	854	32
B1	11	3	543	2249	39	631	16
B1	12	3	648	2771	41	773	16
B1	13	3	291	1196	19	347	5
B1	14	3	606	2548	29	755	28
B1	15	3	506	2267	28	653	17
B2	16	4	408	1930	29	561	25
B2	17	4	383	1774	30	508	22
B2	18	4	596	2661	32	746	27
B2	19	4	555	2665	26	744	34
B2	20	4	564	2695	23	782	28

Next, all readability indices for each text sample were calculated (Table 2); and finally all texts were ordered according to the respective readability indices (Table 3).

Table 2. Readability indices

LL	OD-Code	LL-Code	ARI	C-LI	FRE	F-KGL	GFI	SMOG
A2.1	1	1	1.3	11.6	107.5	0.1	4.8	5.2
A2.1	2	1	2.9	10.4	95.1	2.7	6.9	6.1
A2.1	3	1	4.9	12.9	90.7	3.3	8.1	6.8
A2.1	4	1	1.9	15.0	103.0	0.3	3.9	4.9
A2.1	5	1	4.3	10.7	92.8	3.4	6.2	5.2
A2.2	6	2	3.0	9.1	97.9	2.7	9.0	7.1
A2.2	7	2	2.5	8.2	102.3	2.2	6.3	5.2
A2.2	8	2	5.5	9.5	98.3	3.6	8.3	6.2
A2.2	9	2	4.4	10.0	98.7	2.9	7.9	6.4
A2.2	10	2	7.9	13.4	82.2	5.5	10.8	8.0
B1	11	3	5.0	10.7	94.3	3.5	8.5	6.7
B1	12	3	6.6	11.2	89.8	4.6	8.7	6.6
B1	13	3	5.5	10.3	90.4	4.4	7.8	6.0
B1	14	3	8.8	10.3	80.2	7.2	12.9	8.7
B1	15	3	8.7	12.2	79.3	6.6	10.5	7.5
B2	16	4	7.8	14.1	76.2	6.1	11.7	8.4
B2	17	4	6.7	13.7	81.6	5.0	10.8	8.0
B2	18	4	8.9	12.0	82.0	6.4	11.9	8.3
B2.	19	4	11.8	13.8	71.7	8.5	14.6	9.6
B2	20	4	13.3	13.5	64.6	10.3	14.7	9.4

Table 3. Texts ordered according to readability ease

LL	OD-Code	LL-Code	ARI	C-LI	FRE	F-KGL	GFI	SMOG
A2.1	1	1	1	11	1	1	2	2
A2.1	2	1	4	7	7	4	5	6
A2.1	3	1	8	14	10	7	8	11
A2.1	4	1	2	20	2	2	1	1
A2.1	5	1	6	9	9	8	3	4
A2.2	6	2	5	2	6	5	12	12
A2.2	7	2	3	1	3	3	4	3
A2.2	8	2	11	3	5	10	9	7
A2.2	9	2	7	4	4	6	7	8
A2.2	10	2	15	15	13	14	14	14
B1	11	3	9	8	8	9	10	10
B1	12	3	12	10	12	12	11	9
B1	13	3	10	5	11	11	6	5
B1	14	3	17	6	16	18	18	18
B1	15	3	16	13	17	17	13	13
B2	16	4	14	19	18	15	16	17
B2	17	4	13	17	15	13	15	15
B2	18	4	18	12	14	16	17	16
B2	19	4	19	18	19	19	19	20
B2	20	4	20	16	20	20	20	19

Furthermore, in order to find out significant discrepancies among the readability indices, they were normalized into *z*-scores (Figure 1 and Table 4). A brief examination of the data reveals that the *CL-I* index (*Coleman-Liau*) is the only readability index that exhibits significant deviations compared to the other five ones: in 5 out of 20 texts the deviation of this index exceeded in more than 2 standard deviation measures (texts 1, 4, 7, 14 and 20). Because of this divergence from the rest of the readability indices, we have decided to discard the *Coleman-Liau* readability index for this research.

Figure 1. Z-score normalization of text readability variables

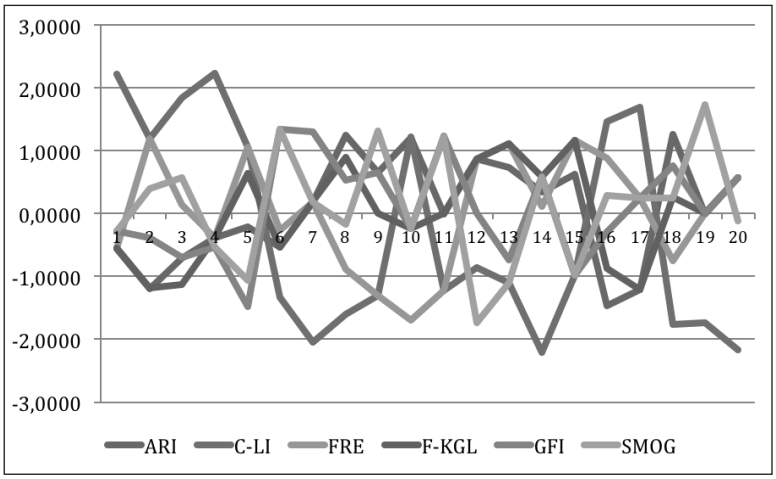

Table 4. Z-score normalization of the ordinal text readability variables

LL	OD-Code	LL-Code	ARI	C-LI	FRE	F-KGL	GFI	SMOG
A2.1	1	1	-0.55	2.21	-0.55	-0.55	-0.27	-0.27
A2.1	2	1	-1.19	1.19	1.19	-1.19	-0.39	0.39
A2.1	3	1	-0.70	1.83	0.14	-1.13	-0.70	0.56
A2.1	4	1	-0.38	2.23	-0.38	-0.38	-0.53	-0.53
A2.1	5	1	-0.21	1.05	1.05	0.63	-1.48	-1.05
A2.2	6	2	-0.53	-1.33	-0.26	-0.53	1.33	1.33
A2.2	7	2	0.18	-2.04	0.18	0.18	1.29	0.18
A2.2	8	2	1.24	-1.59	-0.88	0.88	0.53	-0.17
A2.2	9	2	0.65	-1.30	-1.30	0.00	0.65	1.30
A2.2	10	2	1.21	1.21	-1.69	-0.24	-0.24	-0.24
B1	11	3	0.00	-1.22	-1.22	0.00	1.22	1.22
B1	12	3	0.86	-0.86	0.86	0.86	0.00	-1.73
B1	13	3	0.73	-1.10	1.10	1.10	-0.73	-1.10
B1	14	3	0.34	-2.20	0.11	0.57	0.57	0.57
B1	15	3	0.62	-0.98	1.16	1.16	-0.98	-0.98
B2	16	4	-1.46	1.46	0.87	-0.87	-0.29	0.29
B2	17	4	-1.21	1.69	0.24	-1.21	0.24	0.24
B2	18	4	1.26	-1.76	-0.75	0.25	0.75	0.25
B2	19	4	0.00	-1.73	0.00	0.00	0.00	1.73
B2	20	4	0.56	-2.16	0.56	0.56	0.56	-0.11

Table 5 shows that textbook sample 1 (OD-code =1) is typified by the readability indices with the lowest score (*ARI, FRE* and *F-KGL*) or with the second lowest one (*GFI* and *SMOG*). In contrast, textbook sample 3 is, according to the readability indices, the 8th, 10th, 7th, 8th or 11th highest score. This is a striking case, as this text seems clearly misplaced, though it is placed at the beginning of the A2.1 EFL textbook. According to its RI, this text should not have been placed in the A2.1 book, but in a more advanced level, depending on the readability measures used: pre-intermediate (*ARI, FRE, FKGL* and *GFI*) or even intermediate one (*SMOG*).

Table 5. Text ordered according to readability ease

LL	OD-Code	LL-Code	ARI	FRE	F-KGL	GFI	SMOG	MD
A2.1	1	1	1	1	1	2	2	0.4
A2.1	2	1	4	7	4	5	6	3.2
A2.1	3	1	8	10	7	8	11	5.8
A2.1	4	1	2	2	2	1	1	-2.4
A2.1	5	1	6	9	8	3	4	1
A2.2	6	2	5	6	5	12	12	2
A2.2	7	2	3	3	3	4	3	-3.8
A2.2	8	2	11	5	10	9	7	0.4
A2.2	9	2	7	4	6	7	8	-2.6
A2.2	10	2	15	13	14	14	14	4
B1	11	3	9	8	9	10	10	-1.8
B1	12	3	12	12	12	11	9	-0.8
B1	13	3	10	11	11	6	5	-4.4
B1	14	3	17	16	18	18	18	3.4
B1	15	3	16	17	17	13	13	0.2
B2	16	4	14	18	15	16	17	0
B2	17	4	13	15	13	15	15	-2.8
B2	18	4	18	14	16	17	16	-1.8
B2	19	4	19	19	19	19	20	0.2
B2	20	4	20	20	20	20	19	-0.2

In order to determine the divergences between the textbook placing of the texts and the readability indices, we have calculated the mean divergences (*MD*) of all texts:

$$MD = \left(\frac{\sum RI}{\#RI} \right) - OD\_Code$$

where $\sum RI$ stands for sum of the various readability indices used, *#RI* for the number of readability indices applied and *OD-code* for the reading-order difficulty code within the textbook sequences.

According to the *MD*s, we find four misplaced texts, probably presented to students too early: Text 3: MD 5.8; Text 10: MD 4; Text 14: MD 3.4; and Text 2: MD 3.2.

Similarly, some apparently linguistically less demanding texts are also misplaced, appearing too late in the textbooks: Text 13: MD -4.4; Text 7: MD -3.8; Text 17: MD -2.8; Text 9: MD -2.6; Text 4: MD -2.4.

Data also reveal that some texts seem to have been improperly placed, as their indices are higher/lower for the textbook in which they appear:

- Text 3 – *A2.1*; should be *A2.2*
- Text 14 – *B1*; should be *B2*
- Text 15 – *B1*; should be *B2*
- Text 11 – *B1*; should be *A2.2*
- Text 17 – *B2*; should be *B1*
- Text 13 – *B1*; should be *A2.2*
- Text 7 – *A2.2*; should be *A2.1*

In order to determine the accuracy of the readability indices, we shall first order the texts according to their *Index Means* (*IM*s) and re-typify them as being elementary/A2.1 (*IM*≤5), pre-intermediate/A2.2 (*IM*≥ 5 and ≤10), intermediate/B1 (*IM*≥10 and ≤15) and upper-intermediate/B2 (*IM*≥15). The re-typification (*New LL-Code*) is given in Table 6. *SMOG* and *C-LI* are the least precise ones, although their correlation values are highly significant.

Table 6. Texts re-typified according to IMs

OD-Code	IM	LL	New LL-code
1	1.5	A2.1	1
4	2	A2.1	1
7	2.67	A2.1	1
2	5.33	A2.2	2
3	10	A2.2	2
5	6.5	A2.2	2
6	7.17	A2.2	2
8	7.83	A2.2	2
9	6.5	A2.2	2
11	9.17	A2.2	2
13	8.67	A2.2	2
10	14.5	B1	3
12	11	B1	3
17	14.33	B2	3
14	16.67	B2	4
15	15.17	B2	4
16	16.5	B2	4
18	15.5	B2	4
19	19.33	B2	4
20	19.67	B2	4

Regarding wrong linguistic level assignment, *ARI* and *F-KGL* accounted for five errors; *C-LI* for six errors, although text 13 was two-level wrongly assigned to *B2* instead of *A2.2* (see Table 7); *GFI* for seven errors; *SMOG* for ten errors (and a two-level wrong assignment); and *FRE* for eleven errors.

Table 7. LL-assignment errors

Textbook	New LL-code	ARI	C-LI	FRE	F-KGL	GFI	SMOG
1	1	Correct	Correct	Correct	Correct	Correct	Correct
2	1	Correct	Correct	Correct	Correct	Correct	Correct
3	2	Incor. (1)	Incor. (1)	Incor. (1)	Incor. (1)	Incor. (1)	Incor. (1)
4	1	Correct	Correct	Incor. (1)	Correct	Correct	Incor. (1)
5	2	Correct	Correct	Incor. (1)	Correct	Correct	Correct
6	2	Correct	Correct	Correct	Correct	Incor. (1)	Incor. (1)
7	1	Correct	Correct	Incor. (1)	Correct	Incor. (1)	Incor. (2)
8	2	Incor. (1)	Correct	Incor. (1)	Correct	Incor. (1)	Correct
9	2	Correct	Correct	Correct	Correct	Correct	Correct
10	3	Incor. (1)	Correct	Incor. (1)	Incor. (1)	Incor. (1)	Correct
11	2	Correct	Correct	Incor. (1)	Incor. (1)	Correct	Incor. (1)
12	3	Correct	Correct	Correct	Correct	Correct	Incor. (1)
13	2	Incor. (1)	Incor. (2)	Incor. (1)	Incor. (1)	Incor. (1)	Incor. (1)
14	4	Incor. (1)	Correct	Incor. (1)	Incor. (1)	Incor. (1)	Incor. (1)
15	4	Correct	Incor. (1)	Correct	Correct	Correct	Incor. (1)
16	4	Correct	Incor. (1)	Incor. (1)	Correct	Correct	Correct
17	3	Correct	Incor. (1)	Incor. (1)	Correct	Correct	Incor. (1)
18	4	Correct	Incor. (1)	Correct	Correct	Correct	Correct
19	4	Correct	Correct	Correct	Correct	Correct	Correct
20	4	Correct	Correct	Correct	Correct	Correct	Correct
Total errors		5 (5)	6 (7)	11 (11)	5 (5)	7 (7)	10 (11)

Surprisingly enough, the three readability indices that best adjust to the *New LL-Code* use different measures. As commented on above, *ARI* uses mean word length and mean sentence length, and to obtain the *F-KGL* index, we need mean sentence length and mean syllable per word. On the contrary, *GFI* gets its index from mean sentence length and average number of complex words. In this way, the calculation of the *ARI* and of *CLI* is straightforward; some easy text processing by means of any

standard concordance program will output the information required to calculate this index (e.g. *WordSmith Tools*). Nonetheless, *F-KGL* and *GFI* are more demanding, as we need reliable software syllable counting (i.e. *WordCalc* or *Syllable Counter*). These applications are less consistent and the resulting data might vary significantly.

Regarding complex word count (words with three and more syllables), we performed some preliminary experimenting and evidenced that 95% of all English words with eight or more characters do entail at least three syllables; this is the measure which has been used to calculate the *GFI* index.

4. Modeling a new index

To attempt the modeling of a new readability index able to classify text samples according to reading ease, we shall take:

- The data on the various texts analyzed (Table 1), entailing all the distinct measures required by the individual readability indices examined, and
- The *New LL-Code*, as this is a sort of average measure of all individual readability indices we have considered.

We shall try to model an index by means of *Discriminant Function Analysis* (*DFA*, hereafter). *DFA* is concerned with the problem of assigning individuals, for whom several variables have been measured, to certain groups that have already been identified in the sample. It is used to determine the variables that discriminate between two or more naturally occurring groups (Cantos, 2013). Thus, our aim is not just to measure and model reading ease, but also to look at the dataset that best describes it.

The *DFA*, using all variables (tokens, characters, sentences, syllables and complex words) outputs very promising results: only two errors (see Table 8). One A2.1 text has been assigned to A2.2 (text 2) and a B2 one has been classified as a B1 one (text 15). This gives an overall precision of 90% compared to the best precision scores of two readability indices above (*ARI* and *F-KGL*) of 75%. A further use of *DFA* is that, if it has turned out to be positive, it is possible to generate a predictive discriminant model to classify new cases.

Table 8. Preliminary *DFA*

		Predicted Group Membership				
New LL-code		**A2.1**	**A2.2**	**B1**	**B2**	**Total**
Count	A2.1	3	1	0	0	4
	A2.2	0	7	0	0	7
	B1	0	0	3	0	3
	B2	0	0	1	5	6
%	A2.1	75.0	25.0	0.0	0.0	100.0
	A2.2	0.0	100.0	0.0	0.0	100.0
	B1	0.0	0.0	100.0	0.0	100.0
	B2	0.0	0.0	16.7	83.3	100.0

By means of the *Fisher Coefficients*, we obtain a table (Table 9) with a constant value and a number of coefficients for each of the variables (tokens, characters, sentences, syllables and complex words) with reference to each readability-ease level.

Table 9. Fisher Coefficients

	Readability-ease level			
	A2.1	**A2.2**	**B1**	**B2**
Tokens	-0.14	-0.11	-0.26	-0.26
Characters	-0.06	-0.04	-0.05	-0.05
Sentences	1.36	0.79	1.01	0.63
Syllables	0.35	0.26	0.44	0.46
Complex words	-0.43	-0.20	-0.18	0.00
(Constant)	-21.86	-13.09	-31.81	-31.43

This yields four equations, one for each readability-ease level. To illustrate the potential applicability of the equations above, we can take, for example, a randomly chosen text with tokens = 300; characters = 1,200; sentences = 40; syllables = 400; and complex words = 10, which will be assigned to the readability-ease level with the largest resulting value according to the four functions above. Thus, maximizing the four coefficients we find that this text is most likely to be an A2.1 text, as *Elementary/*

A2.1 is the highest resulting coefficient (44.338); in second place, it would be classified under *Intermediate/B1* (34.239). The least likely group membership would be *Upper-intermediate/B2* (30.672), as the coefficient obtained in the corresponding equation is the lowest one.

5. Conclusions

Readability indices can be useful for the automatic classification of texts, especially within language teaching. Among other applications, they allow for the previous determination of the difficulty level of texts directly extracted from the Internet. The problem is that these readability indices may offer disparity, and this is precisely what has motivated our attempt to unite their potential, utilizing all the variables used by them. A discriminant analysis of all the variables under examination has enabled the creation of a much more precise model, improving the previous best results by 15%. It is also worth noting that errors or disparities in the difficulty level of the analyzed texts have been detected. Specifically, the *DFA* has helped us examine whether the linguistic features contained within the formula were significant predictors of level classification, and what is more, *DFA* has also optimized the predictors by means of re-weighting them (Fisher coefficients), resulting into four new readability indices, one for each LL, with not just new weighting but also a new "combination" of variables.

Our intention is to delve more deeply into the refinement and use of readability indices for tasks such as automatic classification of texts, especially within the area of language teaching, comparing different languages and confirming whether these readability indices offer a similar degree of precision or if they require any adjustment for its calculation as far as variables are concerned.

References

Ash, R. A., & Edgell, S. L. (1975). A note on the readability on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology*, *60*(6), 765–766. http://dx.doi.org/10.1037/0021-9010.60.6.765

Brown, J. D. (1998). An EFL readability index. *Journal of the Japan Association for Language Teaching, 20*(2), 7–36.

Cantos, P. (2013). *Statistical Methods in Language and Linguistic Research*. Sheffield, UK: Equinox Publishing.

Coleman, M., & Liau, T. L. (1975). A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology, 60*(2), 283–284. http://dx.doi.org/10.1037/h0076540

Connaster, B. F. (1999). Last rites for readability formulas in technical communication. *Journal of Technical Writing and Communication, 29*(3): 271–287.

Cronin, M., O'Hanlon, S., & O'Connor, M. (2011). Readability Level of Patient Information Leaflets for Older People. *Irish Journal of Medical Science, 180*(1), 139–142. https://doi.org/10.1007/s11845-010-0624-x

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84–101.

Crossley, S. A., Yang, H. S., & McNamara, D. S. (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language, 26*(1), 92–113.

Cunningham, J. W., & Mesmer, H. A. (2014). Quantitative Measurement of Text Difficulty: What's the Use? *The Elementary School Journal, 115*, 255–269. https://doi.org/10.1086/678292

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*, 37–53.

Dellar, H., & Walkley, A. (2005). *Innovations Elementary. Coursebook*. Boston, MA: Thomson ELT.

___. *Innovations Pre-Intermediate. Coursebook*. Boston, MA: Thomson ELT.

Dellar, H., Walkley, A., & Hocking, D. (2004). *Innovations Intermediate. Coursebook*. Boston, MA: Thomson ELT.

Dellar, H, Hocking, D. & Walkey, A. (2004). *Innovations Upper Intermediate. Coursebook*. (2nd ed.). Boston, MA: Thomson ELT.

DuBay, W. H. (2004). *The Principles of Readability.* Retrieved May 2, 2017, from http://www.impact-information.com/impactinfo

/readability02.pdf

Farr, J. N., Jenkins, J. J., & Paterson, D. G. (1951). Simplification of the Flesch Reading Ease Formula. *Journal of Applied Psychology, 35*(5), 333–357.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford, E. E., & Stenner, A. J. (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology, 107*, 4–29. http://dx.doi.org/10.1037/a0037289

Flesh, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221–233.

Flor, M., & Klebanov, B. B. (2014). Associative Lexical Cohesion as a Factor in Text Complexity. *International Journal of Applied Linguistics: Special issue on Recent Advances in Automatic Readability Assessment and Text Simplification, 165*(2), 223–258. http://dx.doi.org/10.1075/itl.165.2.05flo

Freda, M. C. (2005). The Readability of American Academy of Pediatrics Patient Education Brochures. *Journal of Pediatric Health Care, 19*(3), 151–156. https://doi.org/10.1016/j.pedhc.2005.01.013

Fry, E. (1968). A Readability Formula that Saves Time. *Journal of Reading 11*(7), 513–516.

Greenfield, G. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Doctoral dissertation No. 99–38670. Philadelphia, PA: University Microfilms, Temple University.

Greenfield, J. (2004). Readability formulas for EFL. *Journal of the Japan Association for Language Teaching, 26*, 5–24.

Gunning, R. (1952). *The technique of clear writing*. New York, NY: McGraw-Hill.

Guo, S., Zhang, G., & Zhai, R. (2011). Integrating Readability Index into Twitter Search Engine. *British Journal of Educational Technology, 42*(5), 103–105.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index. Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Memphis, TN: US Naval Air Station.

McLaughlin, G. H. (1969). SMOG grading - A new readability formula. *Journal of Reading, 22*, 639–646.

Maxwell, M. (1978). Readability: Have we gone too far? *Journal of Reading, 21*, 525–530.

Nelson, J, Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. Retrieved from the Council of Chief State School Officers website: https://achievethecore.org/content/upload/nelson_perfetti_liben_measures_of_text_difficulty_research_ela.pdf

Smith, E. A., & Senter, R. J. (1967). *Automated Readability Index*. Retrieved from the Defense Technical Information Center website: http://www.dtic.mil/dtic/tr/fulltext/u2/667273.pdf

Smith, E. A., & Kincaid, P. (1970). Derivation and validation of the Automated Readability Index for use with technical materials. *Human Factors, 12*(1), 457–464.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.

Vajjala, S., & Meurers, D. (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In J. Tetreault (Ed.), *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications* Quebec, Canada: Association for Computational Linguistics. 163-173.

__. (2013). On the Applicability of Readability Models to Web Texts. In S. Williams, A. Siddharthan, & A. Nenkova (Eds.), *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Sofia, Bulgaria: Association for Computational Linguistics. 59-68.

__. (2014a). Assessing the relative reading level of sentence pairs for text simplification. In I. S. Wintner, S. Goldwater, & S. Riezler (Eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* Gothenburg, Sweden: Association for Computational Linguistics. 288-297.

__. (2014b). Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications. *International Journal of Applied Linguistics (Special Issue on Current Research in Readability and Text Simplification), 165*(2), 194–222. https://doi.org/10.1075/itl.165.2.04vaj