ISSN: 2688-9323

Research on the Readability of Reading Materials

Xiaomeng Zhang

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233030, China

Abstract

Aiming at the evaluation of English text reading difficulty, this paper uses TOPSIS comprehensive evaluation method, analytic hierarchy process and other methods to construct the TOPSIS evaluation model after weighting based on analytic hierarchy process, uses Matlab, Python, SPSS and other software to program, and obtains the appropriate difficulty scoring range of English reading materials at different learning stages, which can be used as a reference for the selection of reading materials for language testing with appropriate difficulty level. Firstly, by consulting the relevant literature and combining with the actual life, four first-class evaluation indexes and 10 second-class evaluation indexes are determined. Secondly, from the race data set released by the language technology center of Carnegie Mellon University in 2017, 30 reading materials of primary school, junior middle school, senior high school and College English were obtained and numbered. Python is used to extract the index data, the improved TOPSIS evaluation model with weight analysis is used to evaluate different English texts, and the normalized score results of the difficulty of 120 articles are obtained. According to the scores, the suitable difficulty intervals of different learning stages are divided.

Keywords

Readability of Reading Materials; Hierarchical Analysis; Matlab; TOPSIS.

1. Introduction

As the level of human spiritual civilization rises, people are more widely drawing knowledge from different cultural backgrounds. Meanwhile, with the development of foreign language teaching and the opening up and prosperity of foreign language teaching materials, more and more foreign language articles and reading materials are flooding into the domestic market, and more and more teaching materials are available for students to read and study. These changes have broadened students' knowledge on the one hand, but on the other hand, they have led to some problems. Is the difficulty of the reading material appropriate for the age group? What levels of reading are available for different knowledge bases? Variations in the difficulty and writing quality of materials may have long-term negative consequences for students. Readings that do not match the difficulty and the level of students can undermine students' confidence, affect learning, and lead to half-hearted efforts. English materials should meet the requirements of the curriculum standards in terms of difficulty and weight, as well as the learning goals of teachers and students, and conform to the laws of psychological cognition and development. Therefore, it is especially important to evaluate the difficulty of English reading materials.

2. Literature Review

To this problem, Su [1] proposed in 1997 to analyze the readability of English texts from a linguistic perspective, giving three dimensions for analyzing texts: lexical, semantic, and

ISSN: 2688-9323

syntactic. Tong Zhaojun in 2009 reviewed the two existing evaluation methods dominated by the difficulty coefficient method and the level assessment method, and proposed a method to improve the difficulty evaluation method of college English textbooks by combining word length or word frequency to measure semantic difficulty to derive the reading difficulty value of the text and subjective experience to judge the difficulty of the material. Zhan Xianjun[2] proposed a corpus-based text difficulty assessment for foreign language reading tests in 2014, which provides a multi-dimensional comparison of English language examinations and TOEFL reading in terms of average text length, word-level distribution, morphological character/class character ratio of vocabulary, and sentence structure article topics, providing technical and theoretical insights into the application of corpus technology for pre-test difficulty control and post-test difficulty assessment of foreign language reading tests. This paper addresses this issue, based on new data, and investigates four aspects: lexical category indicators, grammatical category indicators, article cohesion and article description, in an effort to obtain a more scientific and reasonable English reading difficulty evaluation system.

3. Model

3.1. Model Preparation

A special sheet in Common Core State English provides a detailed description of how to determine text difficulty in reading, [3]which introduces the concept of triangular dimensions, namely, the qualitative dimension of text difficulty, the quantitative dimension of text difficulty, and reader and task difficulty:

Figure 1. Text Difficulty 3D Chart

By deeply understanding the meaning of the indicators in the book and analyzing them in relation to the problem, this paper selects four first-level indicators of vocabulary, grammar, text cohesion and text description at the quantitative level to study English reading texts from multiple dimensions.

By reviewing related studies, this paper identifies the indicators of word, sentence, and context, 30 middle school and high school articles each from the RACE dataset released by the Language Technology Center of Carnegie Mellon University in 2017, 30 English reading materials for elementary school and 30 English reading materials for college English level 4 and 6 and examinations on the web, and extracts the word in Microsoft Word The information such as the number of words and paragraphs were processed to get the average number of sentences in paragraphs and the average number of words in sentences. The high-frequency word corpus and all monosyllabic word corpus of different elementary schools, junior high schools, high schools and universities were crawled through Python respectively, and the percentage of high-frequency words and monosyllabic words in the crawled articles were further calculated by the program. The number of occurrences was calculated by iteratively comparing the crawled corpus of pronouns and conjunctions with the collected articles using a similar method, and the raw data for the analysis of this paper were obtained through the above data collection and processing.

Table 1. Selection and classification of each type of index

Tier 1 Indicators	Secondary Level Indicators	Indicator Type
Vocabulary indicators	Number of words	Positive indicators
	Percentage of monosyllabic words	Negative indicators
	Percentage of high-frequency words	Negative indicators
Grammatical indicators	Average number of words in a sentence (number of words, number of sentences)	Positive indicators
	Percentage of single sentences	Negative indicators
Article Cohesiveness	Number of pronouns	Positive indicators
	Number of conjunctions	Negative indicators
	Number of repeated words	Negative indicators
Article Description	Number of paragraphs	Positive indicators
	Average number of sentences in a paragraph	Positive indicators

3.2. Modeling of the Empowered TOPSIS Algorithm

TOPSIS evaluation method is a scientific method commonly used in the multi-objective decision analysis of limited solutions [4], and the use of raw data is more adequate and less information loss. Before establishing the TOPSIS model this paper classifies the index data set up, some data are bigger the better, some data are smaller the better, some data are closer to a certain value the better, some are best in an interval, this different direction and interval makes the analysis confusing, in order to simplify the analysis we forward the data:

Very Small Indicators:

$$x^* = \max(x) - x \tag{1}$$

Intermediate indicators:

$$x^* = 1 - \frac{|x - x_{best}|}{\max(|X - x_{best}|} \tag{2}$$

After forwarding, the data needs to be normalized in order to eliminate the effect of data magnitude:

$$x^* = \frac{x}{\sqrt{\sum_{i=1}^{n} {x_i}^2}} \tag{3}$$

Then the construction of scoring metrics is carried out: the distance to the optimal solution is denoted as D+ and the distance to the worst solution is denoted as D-, and the maximum and minimum values and the maximum and minimum distances are defined:

Define maximum value:

$$k^+ = k_1^+, k_2^+, \cdots, k_m^+ = (\max\{k_{11}, k_{21}, \cdots, k_{n1}\}, \max\{k_{12}, k_{22}, \cdots, k_{n2}\}, \cdots, \max\{k_{1m}, k_{2m}, \cdots, k_{m2}\}) \tag{4}$$

Define minimum value:

$$k^- = k_1^-, k_2^-, \cdots, k_m^- = (\min\{k_{11}, k_{21}, \cdots, k_{n1}\}, \min\{k_{12}, k_{22}, \cdots, k_{n2}\}, \cdots, \min\{k_{1m}, k_{2m}, \cdots, k_{m2}\}) \tag{5}$$

Obtain the final evaluation index S_i:

$$\text{Un-normalized } S_i = D_i^- / (D_i^+ + D_i^-), \text{ after normalization } \widetilde{S_i} = S_i / \sum_{i=1}^{n} S_i$$

where the closer S_i is to 1, the better it is.

3.3. Calculation of Indicator Weights

Such a calculation we found a problem, that is, did not join the weight coefficient, according to the above method to calculate the distance is assumed that each factor for the final evaluation are equivalent, there is no importance, however, in practice this is not possible, so we add weights and then model solution.[5]

Therefore, this paper uses the analytic hierarchy process to weight it, and adds the weight $w_j (j = 1,2,\cdots,10)$ we obtained in the process of defining the distance.

Thus we have:

Define the distance between the first evaluation object and the maximum value

$$D_i^+ = \sqrt{\sum_{j=1}^{10} w_j (k_j^+ - k_{ij})^2} \tag{6}$$

Define the distance between the first evaluation object and the minimum value

$$D_i^- = \sqrt{\sum_{j=1}^{10} w_j (k_j^- - k_{ij})^2} \tag{7}$$

Obtain the final evaluation index $S_i{}^*$:

$$\text{Un-normalized } S_i{}^* = D_i^- / (D_i^+ + D_i^-), \text{ after normalization } \widetilde{S_i{}^*} = S_i{}^* / \sum_{i=1}^{n} S_i{}^*$$

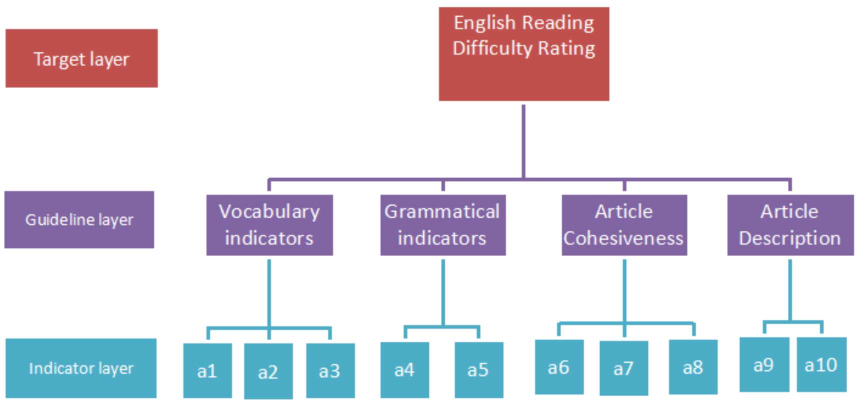

Figure 2. Hierarchical analysis structure chart

where the closer $S_i{}^*$ is to 1, the better it is.

In the analytic hierarchy process,first we need to build a hierarchical structure model: target layer, criterion layer, and solution layer, and we draw a hierarchical analysis structure diagram according to the classification results, as shown in the Figure 2.

After establishing the structural model, we have to start constructing the judgment matrix, we can determine 10 evaluation indicators and compare them at the same level of species indicators, which can get the judgment matrix $A (A = (a_{ij})_{n*n})$. Then the consistency test is performed on the judgment matrix:

$$CI = \frac{\lambda_{max} - n}{n-1} \tag{8}$$

where λ_max is the maximum eigenvalue of the judgment matrix. When CI=CR/RI≤0.10, the consistency of the judgment matrix is considered acceptable; otherwise, certain corrections should be made to the judgment matrix.

4. Conclusion

The importance of each level 1 indicator in the hierarchical analysis method should be on equal footing, so the judgment matrix of English text reading difficulty evaluation indicators was established directly for the 10 level 2 indicators. The pairwise comparison matrix is a comparison of the relative importance of all factors in this level against a factor in the previous level. The element a_{ij} of the pairwise comparison matrix indicates the comparison result of the i factor relative to the j factor, and this value is given using Santy's 1-9 scale method, as shown in the Table 2.

Table 2. Scaling method for the importance of each indicator of hierarchical analysis

Scale	Meaning
1	indicates that the two factors are equally important compared to each other
3	Indicates that one factor is slightly more important than the other when compared to the other factor
5	Indicates that one factor is significantly more important than the other when compared to the two factors
7	Indicates that one factor is strongly more important than the other when compared to the two factors
9	Indicates that one factor is more extremely important than the other when compared to the two factors
2, 4, 6, 8	The median of the above two adjacent judgments
Countdown	Factor i is compared with j to determine a_{ij}, thus $a_{ji} = 1/a_{ij}$

Construct the judgment matrix according to the above scaling method: A=[1 2 0.33 0.5 0.5 2 2 0.5 3 1;0.5 1 0.33 0.5 0.33 2 1 0.33 1 0.5;3 3 3 1 2 1 3 4 1 3 1;2 2 0.5 1 0.5 2 2 0.5 3 1;2 3 1 2 1 4 3 1 2 2 2;0.5 0.5 0.33 0.5 0.5 0.25 1 1 0.33 1 0.5 ;0.5 1 0.25 0.5 0.33 1 1 0.33 1 0.5;2 3 1 2 1 3 3 1 2 2;0.33 1 0.33 0.33 0.5 1 1 0.5 1 0.5 0.5;1 2 1 1 0.5 2 2 0.5 2 1]

Table 3. Specific weights for each indicator

Indicators	Hierarchical analysis weights
Number of words	0.0903
Percentage of monosyllabic words	0.0550
Percentage of high-frequency words	0.1703
Average number of words in a sentence	0.1072
Percentage of single sentences	0.1660
Number of pronouns	0.0465
Number of conjunctions	0.0492
Number of repeated words	0.1614
Number of paragraphs	0.0527
Average number of sentences in a paragraph	0.1015

The CI=0.0267 and CR=0.0179 of the judgment matrix were obtained through Matlab to pass the consistency test, and the eigenvectors corresponding to the maximum eigenvalue λ_max=10.2405 and normalized to obtain the weights of each index.

The entropy weighting method reflects the objective magnitude of the influencing factors through the aggregation of data to derive the weights. The two weights, and the combined weights are listed in the Table 3.

The normalized scores of i=(i=1,2,...,120) evaluation objects can be calculated by the formula. Because it is difficult to list all 120 scores, only the difficulty scores of 5 English reading materials in elementary school, middle school, high school and college are listed in the article.

Table 4. Scoring on the difficulty of English reading materials at different learning levels

Serial number	Reading material number	Score	Serial number	Reading material number	Score
1	Primary_01	0.0788	11	Senior_01	0.5571
2	Primary_02	0.2164	12	Senior_02	0.6398
3	Primary_03	0.1615	13	Senior_03	0.7159
4	Primary_04	0.0992	14	Senior_04	0.7669
5	Primary_05	0.1495	15	Senior_05	0.5946
6	Junior_01	0.2645	16	College_01	0.8803
7	Junior_02	0.2992	17	College_02	0.8093
8	Junior_03	0.5413	18	College_03	0.9465
9	Junior_04	0.3605	19	College_04	0.9877
10	Junior_05	0.3495	20	College_05	0.8572
...	...	...	...	...	...

* Data source: RACE dataset released by Carnegie Mellon University's Center for Language Technology in 2017, reading materials for College English Level 4 and 6, and English for Examinations and Research

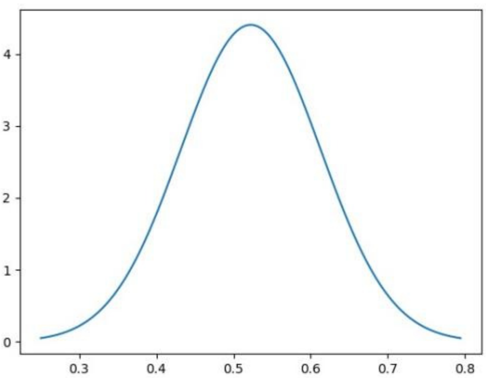

Figure 3. Normal distribution curve of reading difficulty scores

The normalized scores were in the interval of [0,1], the closer the score was to 1 the more difficult the text was to comprehend and suitable for use as test material for reading English in the upper grades, and the closer the score was to 0 the easier the text was to comprehend and suitable for use as test material for reading English in the lower grades. The analysis of the score data yielded four intervals: [0,0.25), [0.25,0.55), [0.55,0.8), and [0.8,1] corresponding to the difficulty scores of reading test materials suitable for elementary school, middle school, high

ISSN: 2688-9323

school, and college, respectively, and then a normal distribution curve with a mean of 0.52 and a variance of 0.09 was fitted to the 120 score data to obtain the Figure 3.

References

[1] Cao Xiaoli. Construction of a linear regression model-based assessment model for the readability of elementary school English textwood [D]. Central University for Nationalities, 2021.

[2] Zheng. Study on the readability of selected texts in elementary school language textbooks[D]. Shanghai Normal University,2017.

[3] Wang Jianxiang. Research and implementation of word vector technology for readability assessment [D]. Nanjing University,2017.

[4] Sun Gang. Research on Chinese text readability prediction method based on linear regression[D]. Nanjing University, 2015.

[5] Wang Maid. Research on the evaluation of readability of general high school English textbooks[D]. Shaanxi Normal University,2015.