Computers, Materials & Continua CMC, vol.61, no.1, pp.213-225, 2019

Readability Assessment of Textbooks in Low Resource Languages

Zhijuan Wang[1, 2], Xiaobin Zhao[1, 2], Wei Song[1, *] and Antai Wang[3]

Abstract: Readability is a fundamental problem in textbooks assessment. For low resources languages (LRL), however, little investigation has been done on the readability of textbook. In this paper, we proposed a readability assessment method for Tibetan textbook (a low resource language). We extract features based on the information that are gotten by Tibetan segmentation and named entity recognition. Then, we calculate the correlation of different features using Pearson Correlation Coefficient and select some feature sets to design the readability formula. Fit detection, F test and T test are applied on these selected features to generate a new readability assessment formula. Experiment shows that this new formula is capable of assessing the readability of Tibetan textbooks.

Keywords: Readability assessment, low resource language, textbook in Tibetan, linear regression, named entity.

1 Introduction

Readability is important for assessing text and it is often used to rate if a reader can read and understand the text easily. The study of readability has a long history, and its research have been widely used in education research, book publishing and online publishing [Dale and Chall (1949)].

There are many methods to assess the readability of textbooks in rich resources languages such as English [Kane (1967)], Chinese [Pang (2006)], French [Uitdenbogerd (2005)], German [Hancke, Vajjala, and Meurers (2012)] and so on. These researched mainly focus on two parts: selecting features and designing the readability assessment model based on selected features. However, there is little research on readability assessment for low resource languages. Here, taking Tibetan, a low resource language, as an example, we conduct an in-depth study on how to assess the readability of Tibetan textbooks. The rest of this paper is organized as follows. In the next section, we talk about the background including features used in readability, the assessment method of readability and readability assessment of Tibetan textbooks. In Section 3, we introduce the corpus we used. In Section 4, we propose the feature selection strategy and a new readability assessment formula for Tibetan Language. The paper concludes in Section 5 with guidance of constructing the readability formula in Tibetan Language and future work.

[1] Minzu University of China, Beijing, China.

[2] National Language Resource Monitoring & Research Center of Minority Languages, Beijing, China.

[3] New Jersey Institute of Technology, 323 Dr Martin Luther King Jr Blvd, Newark, NJ 07102, USA.

[*] Corresponding Author: Wei Song. Email: sw_muc@126.com.

2 Background

The research on readability started in the United States in 1923. Two parts were mainly involved: feature selection and assessment model construction based on features selected.

2.1 Features used in readability assessment

Readability assessment is based on various features. Vogel et al. [Vogel and Washburne (1928)] used the number of words, part of speech, difficult words list and the number of phrases as features to assess the readability of text. Dale et al. [Dale and Chall (1948)] measured the readability by determining the distribution of difficult words in the text through 3000 common vocabularies. Flesch et al. [Flesch (1948)] obtained a readability index from 0 to 100 by calculating the number of syllables per 100 words and the average number of words per sentence. The ATOS for Books [Fry (2000)], developed by an American commercial company, took the length of the text as an important feature in assessing readability. Part-of-speech-based grammatical features were used to assess the readability [Heilman, Collins-Thompson, Callan et al. (2007); Leroy, Helmreich, Cowie et al (2008)]. Feng et al. [Feng, Jansche, Huenerfauth et al. (2010)] thought that the number of named entities in the text will affect the reader's memory burden, and used the number named entity as one of features to measure the readability. Gemoets et al. [Gemoets, Rosemblat, Tse et al (2004)] took personal pronouns as features to measure readability. François et al. [François and Fairon (2012)] used 46 textual features to get the readability of French. Some commonly used features are listed in Tab. 1.

Table 1: The features commonly used in readability assessment

Classes	Features
Word level	NW(The number of words per document), KW(The kinds of words per document), AWL(average word length), Difficult words/easy words list, ASWs(Number of stop words per sentence), ASW(Number of stop words per document), …
Sentence features	ASL(average sentence length per documents), ANS(average number of sentences per documents), …
POS features	ANN(number of nouns per documents), AJJ(number of adjectives per documents), ANNP(number of proper nouns per documents), ARB(Number of adverbs per document), ACC(Number of coordinating conjunctions per document), APRP(Number of personal pronouns per document), …
Others features	NNE(Number of named entities per document), ASe(Number of senses per word per document), ARP(Number of particles per document), ASym(Number of symbols per document), FSL(Features specific to language), …

2.2 Assessment method

There are some methods to assess the readability, which can be divided into two categories: formula and other methods.

2.2.1 Readability assessment Formulas

Two American, Lefley and Presie [Jia (2015)], designed the first readability assessment formula. More than one hundred readability formulas are produced. But only seven of them are commonly used [Tekfi (1987)]. Here, we just introduce these seven formulas briefly.

1. Vogel & Washbune [Vogel and Washburne (1928)]

This formula was first synthesized by Vogel and Washburne in 1928. The Vogel & Washbune formula is as follows:

$$Y = 17.43 + 0.085 * KW + 0.101 * X_1 + 0.604 * X_2 - 0.411 * NW \quad (1)$$

where, X_1 is the number of prepositions, X_2 is the number of complex words (words with more than three syllables)

2. Flesch Reading Ease [Flesch (1948)]

This formula was designed by Flesch in 1948. He graded the score according to the formula and proposed a range of legibility of 0-100. The Flesch Reading Ease formula is as follows:

$$Y=206.835-1.015*ASL-84.6*ASW \quad (2)$$

3. Gunning Fog Index [Gunning (1969)]

This formula was created by American professor Robert Gunning in 1952. The lower the Fog index of the article, the easier it is for readers to understand. The Gunning Fog Index formula is as follows:

$$Y=0.4\ (ASL+PHW) \quad (3)$$

where, PHW is the percentage of hard words.

4. Automated Readability Index (ARI) [Senter and Smith (1967)]

This formula was proposed by Senter and Smith in 1967. It is based on linear regression analysis. The lower the ARI index of the article, the easier it is for readers to understand. The Automated Readability Index formula is as follows:

$$ARI=4.71*AWL+0.5*ASL-21.43 \quad (4)$$

5. Flesch-Kincaid Formula (FK) [Kincaid, Fishburne, Rogers et al. (1975)]

This formula was jointly designed by Kincaid and Flesch in 1975. It is the US Department of Defense's standard readability formula and is also a built-in readability formula for Microsoft Office. The Flesch-Kincaid Formula is as follows:

$$Y=0.39\times ASL+11.8\times AWL-15.59 \quad (5)$$

6. SMOG Grading [Mc Laughlin (1969)]

This formula was constructed by G. Harry McLaughlin in 1969 and it is the only formula that has only one feature. The lower the SMOG index of the article, the easier it is for readers to understand. The SMOG grading formula is as follows:

$$Y = 3 + \sqrt{X} \tag{6}$$

Where, X is the number of multi-syllable words in 30 sentences.

7. Dale-Chall [Dale and Chall (1948)]

This formula was originally designed by Dale and Chall in 1948 and was revised in 1995. The formula is based on a common vocabulary that has been expanded from the original 763 common words to 3,000 common words. The Dale-Chall formula is as follows:

$$Y = (0.1579 \times X_1) + (0.0496 \times X_2) + 3.6365 \tag{7}$$

where, X_1 is the percentage of uncommon words (based on Dale 3000 common vocabulary), and X_2 is the average number of words per sentence.

To construct readability formula, different methods (such as logistic regression, linear regression) are used to construct readability formulas based on selected features.

2.2.2 Other methods

Besides formula, there are some other methods to assess the readability of texts.

The cloze method was first proposed by Taylor [Taylor (1953)]. In the "Taylor cloze test", the word after every fifteen words in an article was deleted, and then students of different grade groups were asked to fill the deletion. The word, when the correct rate of a group of answers exceeds 50%, classifies the readability of the article into the group level that can be read easily.

The subjective assessment method invites experts and scholars in related fields to judge the difficulty of the text by artificially determining the readability of the text. In the absence of highly knowledgeable experts and scholars, multiple questionnaires are distributed to teachers or students. Finally, judge if the texts can be read easily or not. The main advantage of this method is that it is simple and easy [Jia (2015)].

In recent years, with the development of machine learning, the readability assessment method based on machine learning has also began to be explored. Chen et al. [Chen, Tsai and Chen (2011)] used TF-IDF and SVM to assess Chinese readability. François et al. [François and Fairon (2012)] combined methods with knowledge of machine learning to study the readability. Hancke [Hancke (2012)] and Vajjala et al. [Vajjala and Meurers (2012)] transformed the problem of readability into a classification problem, and used text classification to evaluate the readability of text.

2.3 Readability assessment of Tibetan textbooks

At present, there is little research on readability of Tibetan textbooks. Most researches are mainly focus on the vocabulary statistics of Tibetan textbooks. For example, Cao et al. [Cao, Han and Dong (2012)] measured the vocabulary of junior high school and high school Tibetan textbooks which are compiled by the five provinces (districts) Tibetan Language Compilation Committee. Zhang et al. [Zhang, Gao, Li et al. (2010)] used statistical methods to analyze the article genre, literary genre and material selection of the new and old versions of the primary school Tibetan textbooks. Wang et al. [Wang (2012)] used the junior middle school Tibetan textbook as a corpus to conduct a shallow syntactic

analysis and proposed six Tibetan block types. Renqing Zhuoma et al. [Renqing Zhuoma (2015)] studied the mistranslation of the translation in the textbook.

As described above, we can see little research has been carried out on readability of Tibetan textbooks which make it necessary to study the readability of Tibetan textbooks.

3 Corpus

Tibetan is a low resource language which is a cluster of Sino-Tibetan languages and spoken primarily by Tibetan peoples, who live across a wide area of eastern Central Asia. As an alphabetic writing language, Tibetan has 30 consonants and 4 vowel signs. Its smallest grammar unit is syllable. "་" is the mark of syllable. One or more alphabets compose a syllable and one or more syllables can compose a word.

Fig. 1 is an example of Tibetan sentence. A Tibetan syllable includes root letter, prefix, head letter, vowel, suffix and post suffix. There is no white space between Tibetan words.

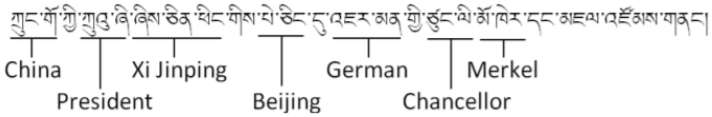

Figure 1: An example of Tibetan sentence

As shown in Tab. 1, many features used in assessing the readability of textbook are based on words. So, Tibetan texts should be segmented firstly. We use Tibetan word segmentation software (developed by the Institute of Ethnology and Anthropology of the Chinese Academy of Social Sciences [Li, Liu, Long et al. (2018)]) to segment the Tibetan words.

Here, we use Tibetan primary textbook as assessment corpus, which is created by Tibetan Language Teaching Materials Compilation Committee of five provinces (districts) and published by Qinghai Nationalities Publishing press. There are 11 volumes, 261 articles and 5, 5198 words. Tab. 2 shows the basic information of this corpus.

4 Readability assessment formula for Tibetan textbooks

At present, the performance of part-of-speech system of Tibetan is not good enough to measure the readability of Tibetan textbooks. Therefore, using Tibetan segmentation system [Li, Liu, Long et al. (2018)] and named entity recognition system [Liu and Wang (2017)], eight features of Tibetan textbooks are extracted. They are:

- NW: the number of words per documents
- KW: the kinds of words per documents
- AWL: average word length per documents
- NNW: the number of new words per documents.
- ASL: average sentence length per documents
- ANS: the number of sentences per documents
- NPRP: number of personal pronouns per document
- NNE: the number of named entity per documents

Table 2: The information of Tibetan textbooks

Volume	The number of words	The kinds of words	The number of documents
2	1017	673	24
3	1803	826	22
4	2492	1012	24
5	3530	1874	24
6	3735	1925	24
7	4878	2178	24
8	1841	2043	26
9	6670	3316	24
10	8340	3529	24
11	11561	4217	24
12	9331	3818	21
Total	55198	25411	261

4.1 Frame

In this paper, linear regression is used to construct the readability assessment formula for Tibetan textbooks. Linear regression formula is shown in Eq. (8).

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \quad (8)$$

where, y is the function, β_0 is the regression constant, X_i is variable, β_i is the regression coefficient of X_i, and ε is the random error.

Fig. 2 introduces the frame of readability assessment formula of Tibetan textbooks. It includes two parts: feature selection and formula construction.

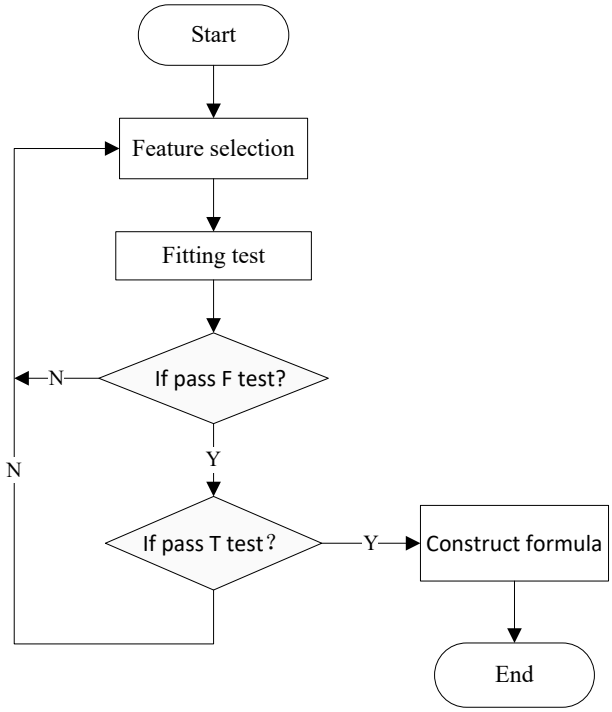

Figure 2: The frame of readability assessment formula in Tibetan

- Features selection. In multi-linear regression model, collinear problems often occur among different variables, which means the changing tendencies of two or more variables are in a same direction. It will weaken the accuracy and stability of the parameter estimation of linear regression analysis [Yang (2004)]. So, if two features are related, only one can be reserved while the others should be moved. Then, we will get several feature sets.
- Formula construction. Three tests are used to construct formula. Fitting test reflect the relation of function and features sets. If the value of fitting test is higher (less than 1), it means the features set can express function better. So, the feature set which has the highest fitting value should be selected. According to the SPSS regression analysis of Christian, when the significance of F test is less than 0.05, it indicates that at least one feature can effectively predict the function and this feature or these features pass the F test. The T test is a test of the regression coefficients for all features of the regression analysis. If its significance is bigger the 0.05, it means that some feature is not pass the T test and it should be removed. Repeat this process until all features pass fitting test, F test and T test. Then, readability assessment formula is gotten.

4.2 Features selection

In order to select features, Pearson Correlation Coefficient is used to calculate the correlation of different features. Its equation is shown in Eq. (9).

$$\text{Pearson}(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{(\sum x^2 - \frac{(\sum x)^2}{N})(\sum y^2 - \frac{(\sum y)^2}{N})}} \quad (8)$$

where (x, y) is the specified pair of variables and N is the total number of variables.

Tab. 3 shows the Pearson correlation of eight features based on the calculation on our Tibetan corpus.

Table 1: Correlation analysis of affecting factors based on Pearson

Features	NW	KW	AWL	NNW	ASL	NS	NPRP	NNE
NW	1	-.63	.40	-.67	.74	.97	.38	.97
KW	-.64	1	-.35	.78	-.66	-.60	-.33	-.66
AWL	.40	-.35	1	-.76	.61	.33	.69	.29
NNW	-.67	.78	-.76	1	-.83	-.61	-.77	-.66
ASL	.74	-.66	.61	-.83	1	.59	.71	.78
NS	.97	-.60	.33	-.61	.60	1	.25	.93
NPRP	.37	-.33	.69	-.77	.71	.25	1	.39
NNE	.97	-.66	.29	-.66	.78	.93	.39	1

If Pearson correlation is high (bigger than 0.5), it means there is a certain correlation between two features and these two features should not exist at the same time. Here, the Pearson correlation of NW (the total number of words) and KW (the kinds of words) is 0.643 which means that there is a certain correlation between them. So, these two features cannot exist at the same time. Here, NW is reserved. Using this as a basis for feature selection, five sets of features are selected as shown in Tab. 4.

Table 2: Five sets of features

No.	Features
1	NW, AWL, NPRP
2	AWL, NW, NS, NNE
3	NS, AWL, NPRP
4	NPRP, NW, NS, NNE
5	NNE, AWL, NPRP

4.3 Construction of Readability assessment formula

We choose SPSS as linear regression analysis tool. Tab. 5 shows the results of the fitting test of the five sets of features. It is clear that the fitness rankings are: set 2>set 1>set 3>set5>set 4. Therefore, set 2 (AWL, NW, ANS, NNE) is selected.

Table 3: Fitting test

No.	Features	R	R^2
1	NW, AWL, NPRP	.995	.990
2	AWL, NW, ANS, NNE	.993	.986
3	ANS, AWL, NPRP	.997	.994
4	NPRP, NW, ANS, NNE	1.000	1.000
5	NNE, AWL, NPRP	.998	.996

Tab. 6 shows the F test of set 2. Clearly, it passes the F test as its significance is less than 0.05.

Table 4: F test

	Sum of square	Degree of freedom	Mean square error	F	Significance
Regression	47.174	5	9.435	60.632	.000
Residual	0.778	5	1.56		
Total	47.952	10			

Tab. 7 shows the T test of regression coefficient of set 2 features. The significance of constants, AWL, ANS and NNE are less than 0.05, thus they pass T test. The significance of NW, however, is bigger than 0.05. Therefore, it is necessary to remove the NW.

Table 5: T test

Feature	Non-standardized coefficient		T	Significance
	B	Standard error		
Constant	-29.385	5.239	-5.609	.002
AWL	24.237	2.894	8.374	.000
NW	17.932	16.203	1.107	.319
ANS	-.391	.086	-4.525	.006
NNE	1.615	.404	4.000	.010

Because the feature set has been changed, linear regression analysis should be performed again until all features pass fitting test, F test and T test. Tabs. 8 to 10 show the second fitting test, F test and T test respectively. It is obvious that all the features pass the fitting test, F test and T test.

Table 6: The second fitting test

No.	Features	R	R^2
2	AWL, NW, ANS, NNE	.988	.977

Table 7: The second F test

	Sum of square	Degree of freedom	Mean square error	F	Significance
Regression	46.832	3	15.611	97.621	.000
Residual	1.119	7	.160		
Total	47.952	10			

Table 8: The second T test

Feature	Non-standardized coefficient		T	Significance
	B	Standard error		
Constant	-34.671	3.861	-8.981	.000
AWL	26.629	2.325	11.454	.000
ANS	-.321	.054	-5.928	.001
NNE	2.057	.235	8.763	.000

According to the linear regression model analysis, Eq. (10) is the readability assessment formulas of the Tibetan textbooks.

$$Y=-34.67+26.629*AWL-0.321*ANS+2.057*NNE \quad (10)$$

4.4 Evaluation of readability assessment formula of Tibetan textbooks

Fig. 3 is readability of Tibetan textbooks based on Eq. (10). From this figure, we can see that, except for volume 7, from volume 2 to volume 10, the value of readability is gradually increasing while the value of readability of volume 7 is increased sharply. The value of readability of volume 8 is decreased and the value of readability of Volumes 11 and 12 is lower than Volumes 10, and their changes are very small.

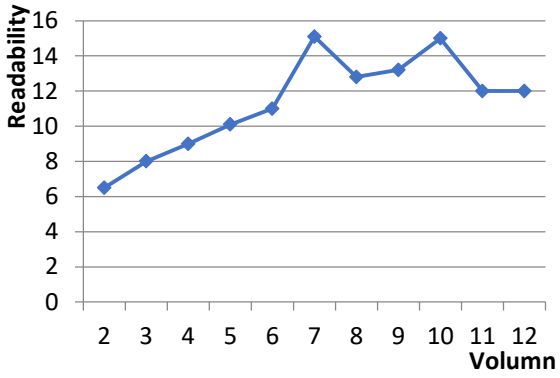

Figure 3: Readability of Tibetan textbooks based on Eq. (10)

5 Conclusion

Formula is the one of most commonly used methods in evaluating the readability of texts. Little research has been carried out on Tibetan readability assessment. We extract eight features using Tibetan NLP tools, and select three features (AWL (average word length), ANS (average number of sentences per documents) and NNE (Number of named entities per document)) to construct readability formula. Then the new formula is constructed based on fitting test, F test and T test. The new formula has good performance and is able to be applied to assess the readability of Tibetan textbooks.

In the future, we will do more research on the methods of feature selection. Also, we will try to use other machine learning model to assess the readability of low resource languages.

Acknowledgements: This work was supported by the China National Natural Science Foundation No. (61331013) and the Young faculty scientific research ability promotion program of Minzu University of China.

References

Cao, H.; Han, X. B.; Dong, X. F. (2012): Statistical analysis of vocabulary measurement of Tibetan textbooks in middle schools. *Science and Technology*, vol. 20, pp. 187-189.

Chen, Y. H.; Tsai, Y. H.; Chen, Y. T. (2011): Chinese readability assessment using TF-IDF and SVM. *Machine Learning and Cybernetics*, vol. 2, pp. 705-710.

Dale, E.; Chall, J. S. (1948): A formula for predicting readability: instructions. *Educational Research Bulletin*, pp. 37-54.

Dale, E.; Chall, J. S. (1949): The concept of readability. *Elementary English*, vol. 26, no. 1, pp. 19-26.

Feng, L.; Jansche, M.; Huenerfauth, M.; Elhadad, N. (2010): A comparison of features for automatic readability assessment. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 276-284.

Flesch, R. (1948): A new readability yardstick. *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221.

François, T.; Fairon, C. (2012): An AI readability formula for French as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 466-477.

Fry, E. (2002). Readability versus leveling. *Reading Teacher*, vol. 56, no. 3, pp. 286-291.

Gemoets, D.; Rosemblat, G.; Tse, T.; Logan, R. A. (2004): Assessing readability of consumer health information: an exploratory study. *Medinfo*, pp. 869-873.

Gunning, R. (1969): The fog index after twenty years. *Journal of Business Communication*, vol. 6, no. 2, pp. 3-13.

Hancke, J.; Vajjala, S.; Meurers, D. (2012): Readability classification for German using lexical, syntactic, and morphological features. *Proceedings of COLING 2012*, pp. 1063-1080.

Heilman, M.; Collins-Thompson, K.; Callan, J.; Eskenazi, M. (2007): Combining lexical and grammatical features to improve readability measures for first and second language texts. *Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 460-467.

Islam, Z.; Mehler, A.; Rahman, R. (2012): Text readability classification of textbooks of a low-resource language. *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pp. 545-553.

Jia, X. X. (2015): A Survey on readability. *Overseas English*, vol. 4, pp. 165-166.

Kane, R. B. (1967): The readability of mathematical English. *Journal of Research in Science Teaching*, vol. 5, no. 3, pp. 296-298.

Kincaid, J. P.; Fishburne Jr, 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 (research report). *University of Central Florida.*

Leroy, G.; Helmreich, S.; Cowie, J. R.; Miller, T.; Zheng, W. (2008): Evaluating online health information: beyond readability formulas. *AMIA Annual Symposium Proceedings*, vol. 2008, pp. 394.

Li, B. H; Liu, H. D; Long, C. J; Wu, J. (2018): Tibetan word segmentation method based on deep learning. *Computer Engineering and Design*, vol. 1, pp. 194-198.

Liu, F. F.; Wang, Z. J. (2017): The study of graininess for Tibetan named entity recognition. *4th Annual International Conference on Information Technology and Applications*, vol. 12, pp. 56-61.

Mc Laughlin, G. H. (1969): SMOG grading-a new readability formula. *Journal of Reading*, vol. 12, no. 8, pp. 639-646.

Pang, L. T. (2006): *Chinese Readability Analysis and Its Applications on the Internet. (Ph. D. Thesis).* The Chinese University of Hong Kong, Hong Kong.

Ren, Q. Z. (2015): *On Relevant Issues in the Translation of Tibetan Chinese Textbooks in Middle Schools (Ph. D. Thesis)*. Tibet University, Tibetan.

Smith E. A.; Senter R. J. *(1967):* Automated readability index. *Competitor New York,* pp. 1-14.

Taylor, W. L. (1953): "Cloze procedure": a new tool for measuring readability. *Journalism Bulletin*, vol. 30, no. 4, pp. 415-433.

Tekfi, C. (1987): Readability formulas: an overview. *Journal of Documentation*, vol. 43, no. 3, pp. 261-273.

Uitdenbogerd, A. (2005): Readability of French as a foreign language and its uses. *ADCS 2005: The Tenth Australasian Document Computing Symposium*, pp. 19-25.

Vajjala, S.; Meurers, D. (2012): On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics*, pp. 163-173.

Vogel, M.; Washburne, C. (1928): An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, vol. 28, no. 5, pp. 373-381.

Wang, W. L. (2012): *A Shallow Syntactic Analysis of Junior Middle School Tibetan Language Teaching Materials (Ph. D. Thesis)*. Northwest University for Nationalities, Gansu.

Xiong, Z. Y.; Shen, Q. Q.; Wang, Y. J.; Zhu, C. Y. (2018): Paragraph vector representation based on word to vector and CNN learning. *Computers, Materials & Continua*, vol. 55, no. 2, pp. 213-227.

Yang, D. H. (2004): *An Example of the Application of SPSS Software in Language Research*. China Social Sciences Press, China.

Zhang, J. S.; Gao, L.; Li, Y. H.; Yu, H. Z. (2010): Quantitative and contrastive analysis of Tibetan and Chinese textbooks in new and old primary schools-statistical research on article genre, literary genre and material selection. *Journal of Northwest University for Nationalities: Natural Science Edition*, vol. 1, pp. 87-91.