Assessing the Readability of Literary Texts in Vietnamese Textbooks

An-Vinh Luong*, Diep Nguyen†, Dien Dinh‡
*University of Science, VNU-HCM, Vietnam. Email: anvinhluong@gmail.com
†University of Social Sciences and Humanities, VNU-HCM, Vietnam. Email: nhudiep2004@gmail.com
*University of Science, VNU-HCM, Vietnam. Email: ddien@fit.hcmus.edu.vn

***Abstract*—Text readability has a huge impact on the reading ability and fully understand the text. Studies on the difficulty of the text have long been concerned in English and some other popular languages. In Vietnamese, there are very few studies on the degree of difficulty of the text, and most of them have been done more than three decades ago. This paper presents a new method for assessing the Readability of Literary Texts in Vietnamese Textbooks based on some specific features of Vietnamese language. The experimental results show that proposed method remarkably improves the accuracy of the assessing.**

***Index Terms*—Vietnamese text readability, Text difficulty, School textbooks**

I. Introduction

Text readability – as defined by Brown et al. [1] – is "a concept that describes the degree to which a text is easy or difficult to read. A readability index is a numerical scale that estimates the readability or degree reading difficulty that native speakers are likely to have in reading a particular text". Text readability has a huge impact on the reading and comprehending a text. Base on the readability, readers can determine whether a text is suitable for their reading ability or not. The text author(s) can also use the readability of the draft to guide readers object or have some adjustments to make it fit the toward reader.

Building a model to analyze text readability supports a lot of practical benefits such as helping scientists to write more readable research reports; supporting educators to draft appropriate textbooks and curricula for each of learners' age; designing publishers to shape their own audiences; helping governments to draft legal documents to adapt the majority of citizens' literacy; assisting manufacturers in preparing user guide for their products, or helping teachers to select curriculum effectively for native learners and even foreigners. Thus, text readability is useful for various fields in society.

Much researches have been done for other languages like Arabic, Italian, French, Chinese, Japanese and so on, but English is still the dominating language in this field. Most famous publications in text readability are about creating linear functions to assess and grade documents, for instance, the Dale-Chall formula [2], the Flesch Reading Ease formula [3], the Flesch-Kincaid formula [4], the Gunning Fog formula [5], the SMOG formula [6], *etc*.

Several years ago, some machine learning approaches have been examined for assessing the text readability like the works of Si and Callan [7] in 2001, Schwarm and Ostendorf [8] in 2005, Heilman et al. [9] in 2007, Tanaka-Ishii et al. [10] in 2010, Vajjala and Meurers [11] in 2012, *etc*.

In Vietnamese, there are few studies on text readability, such as the work of Liem and Henkin [12], [13]. In 2017, when examining the features use for assessing texts in literary textbooks for Vietnamese students using SVM classifier, Luong et al. [14] pointed out that the text length features are very valuable for the classification, but the results are still limited. The need of further examining for higher accuracy when assessing Vietnamese text readability is necessary. In this paper, we propose a new method for assessing the Readability of Literary Texts in Vietnamese Textbooks based on Sino-Vietnamese Words. The remaining of the paper is organized as follows: Section II describe the corpus we used for experiments; Section III lists the features we used for examining; after that, Section IV describes our experiments and results; finally, Section V is our discussions and conclusions.

II. Corpus

We use documents extracted from the textbooks of Vietnamese for elementary students and Literature for middle and high school students as experimental corpus. In Vietnam, the primary school is divided into

978-1-5386-7983-8/18/$31.00 ©2018 IEEE

TABLE I
STATISTICAL NUMBERS OF EACH GRADE.

	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12
No. of documents	67	62	40	40	28	13	17	21	15	19	49
Avg. no. of sentences	18.3	19.6	21.5	21.4	54.8	46.4	65.8	107.3	60.7	105.2	111.7
Avg. no. of words	158	192	231	244	680	677	969	1447	862	1360	1710
Avg. no. of syllables	178	222	276	288	784	821	1131	1710	1006	1579	2179
Avg. no. of characters	827	1065	1335	1396	3709	3942	5402	8160	4860	7535	10761
Avg. no. of distinct words	100.6	125.6	144.3	152.8	304.9	329.7	394.3	526.3	368.4	510.0	576.0
Avg. no. of distinct syllables	111.4	141.5	164.8	173.4	327.5	372.5	428.4	555.5	390.1	534.9	594.2
Avg. sentence length calculated by words	9.1	10.6	11.6	12.7	14.0	18.0	17.8	18.2	15.0	15.7	16.7
Avg. sentence length calculated by syllables	10.4	12.3	14.1	15.2	16.1	22.3	21.3	22.1	17.7	18.7	22.2
Avg. sentence length calculated by characters	48.3	59.6	68.6	74.4	76.7	108.7	103.4	106.6	85.8	90.7	111.2
Avg. word length calculated by syllables	1.1	1.2	1.2	1.2	1.1	1.2	1.2	1.2	1.2	1.2	1.3
Avg. word length calculated by characters	5.2	5.6	5.8	5.8	5.4	6.0	5.7	5.8	5.7	5.7	6.6
Avg. no. of dialect words	13.5	12.3	14.7	13.3	42.2	34.2	48.7	80.1	43.1	70.2	80.3
Avg. no. of distinct dialect words	7.0	7.9	9.4	9.0	19.4	18.2	23.4	29.7	19.1	29.1	26.5
Avg. no. of Sino-Vietnamese words	32.3	44.3	55.2	56.8	149.8	163.6	226.8	337.0	212.3	325.6	497.1
Avg. no. of distinct Sino-Vietnamese words	21.4	31.0	37.3	40.1	81.1	101.8	117.2	158.9	116.7	158.0	214.9
Avg. no. of proper nouns	5.16	6.45	5.83	7.55	18.32	14.69	13.00	32.43	40.47	36.16	59.37
Avg. no. of distinct proper nouns	1.97	3.11	3.28	4.35	7.00	6.15	7.29	13.29	17.53	15.58	22.92

TABLE II
STATISTICAL NUMBERS OF EACH GROUP-OF-2-GRADE.

	Grade 2–3	Grade 4–5	Grade 6–7	Grade 8–9	Grade 10–11	Grade 12
No. of documents	129	80	41	38	34	49
Avg. no. of sentences	18.96	21.48	52.10	88.74	85.53	111.65
Avg. no. of words	175	238	679	1234	1140	1710
Avg. no. of syllables	199	282	796	1451	1326	2179
Avg. no. of characters	942	1365	3783	6926	6355	10761
Avg. no. of distinct words	113	149	313	467	448	576
Avg. no. of distinct syllables	126	169	342	499	471	594
Avg. sentence length calculated by words	9.84	12.14	15.27	18.03	15.39	16.68
Avg. sentence length calculated by syllables	11.31	14.65	18.09	21.74	18.28	22.17
Avg. sentence length calculated by characters	53.72	71.49	86.82	105.17	88.51	111.21
Avg. word length calculated by syllables	1.15	1.20	1.17	1.20	1.18	1.32
Avg. word length calculated by characters	5.42	5.81	5.60	5.76	5.69	6.59
Avg. no. of dialect words	12.93	13.96	39.66	66.05	58.26	80.35
Avg. no. of distinct dialect words	7.40	9.18	18.98	26.87	24.68	26.47
Avg. no. of Sino-Vietnamese words	38.07	56.03	154.17	287.68	275.59	497.08
Avg. no. of distinct Sino-Vietnamese words	26.02	38.70	87.68	140.24	139.76	214.90
Avg. no. of proper nouns	5.78	6.69	17.17	23.74	38.06	59.37
Avg. no. of distinct proper nouns	2.52	3.81	6.73	10.61	16.44	22.92

five school years — from grade 1 to grade 5. However, Vietnamese language textbooks for grade 1 are only exercises for reading simple letters and words, so we did not collect grade 1 textbooks. At the middle school, the course is divided into four academic years — from grades 6 through 9. At the high school, the course is divided into three academic years — from grades 10 through 12.

Because there is no digital resource of these textbooks, documents need to be collected manually by OCR and then post edited them by the following task: (i) Spelling correction; (ii) Punctuation standardized; (iii) Encoding standardized; (iv) Sentence segmentation; (v) and word segmentation. The final corpus contain 371 documents from Grade 2 to Grade 12. For examining, we group the collected documents by 3 levels: (i) by school(Primary School; Middle School and High School); (ii) by 2 continuous grade (2–3, 4–5, 6–7, 8–9, 10–11 and 12); and (iii) by each grade. Tables I, II and III present the statistical numbers of the corpus.

III. FEATURES

In this section, we will describe some features that are used in the study of Luong et al. [14] along with other features we examined.

Average sentence length: this is an the simple and common features when examining the readability of the text. In this study, we used the Average Sentence Length calculated by Words (**ASLW**), by Syllables (**ASLS**) and by Characters (**ASLC**).

Average word length: in this research, we used the Average Word Length calculated by Syllables (**AWLS**) and by Characters (**AWLC**) for examined.

Percentage of difficult words (PDW): in many studies, the percentage of difficult words is one of the most value features when evaluating text readability. However, creating the difficult word list costs a lot

TABLE III
STATISTICAL NUMBERS OF EACH SCHOOL.

	Primary school	Middle school	High school
No. of documents	209	79	83
Avg. no. of sentences	19.92	69.72	100.95
Avg. no. of words	199	946	1477
Avg. no. of syllables	231	1111	1830
Avg. no. of characters	1104	5295	8956
Avg. no. of distinct words	126	387	523
Avg. no. of distinct syllables	142	417	544
Avg. sentence length calculated by words	10.72	16.60	16.15
Avg. sentence length calculated by syllables	12.59	19.85	20.58
Avg. sentence length calculated by characters	60.52	95.65	101.91
Avg. word length calculated by syllables	1.17	1.19	1.26
Avg. word length calculated by characters	5.57	5.68	6.22
Avg. no. of dialect words	13.33	52.35	71.30
Avg. no. of distinct dialect words	8.08	22.77	25.73
Avg. no. of Sino-Vietnamese words	44.94	218.39	406.35
Avg. no. of distinct Sino-Vietnamese words	30.87	112.96	184.12
Avg. no. of proper nouns	6.13	20.33	50.64
Avg. no. of distinct proper nouns	3.01	8.59	20.27

for examining and evaluation, so most researches used frequent word list as a replacement: The more frequent the word is, the easier it is. In this study, we used the top 3,000 frequent words, as the easy word list, extracted from the frequent word list of Dien and Hao [15]. Therefore, the words which are not in this list is considered as difficult words. Similarly, the **Percentage of Difficult Syllables (PDS)** was examined using the top 3,000 frequent syllables extracted from the list of Dien and Hao [15]: the syllables which are not in this list is considered as difficult syllables.

Text length: According to Luong et al. [14], the text length features plays an important role in assessing the text readability for documents in textbooks, because in the time of a lesson, students will only be able to read the texts with appropriated lengths. Therefore, in this study, we also used the text length features proposed by Luong et al. [14]: total number of Sentences (**NSen**), Words (**NWo**), Syllables (**NSyl**), Characters (**NCha**), total number of Distinct Words (**NDWo**) and total number of Distinct Syllables (**NDSyl**).

Our additional features: In this study, we used some additional features for assessing the Vietnamese text readability. They are Part-of-Speech features and specific features for Vietnamese language.

- **Percentage of Sino-Vietnamese words (PSVW):** Vietnamese culture has been deeply influenced by Chinese civilization. Vietnamese language is similar, more than 60% of the Vietnamese words are Chinese origin — these words are called as "từ Hán-Việt" 'Sino-Vietnamese words'. Sino-Vietnamese words often appear in scientific, technical, official and other formal texts, so they are considered more difficult than pure Vietnamese words. Therefore, the percentage of Sino-Vietnamese words in the text maybe valuable for assessing Vietnamese text readability. In addition, the **Percentage of distinct Sino-Vietnamese words (PDSVW)** is also a feature we examined.
- **Percentage of dialect words (PDiaW):** The

TABLE IV
CLASSIFICATION RESULTS PERFORMED ON THE GRADE-BY-GRADE DOCUMENTS.

FEATURES	ACCURACY
Baseline (NSen, NWo, NCha, ASLS, AWLC, PDS, PDW)	0.4446
NSen, NWo, NCha, ASLS, AWLC, PDS, PDW, PSVW, PDSVW	**0.4555**
NSen, NWo, NCha, ASLS, AWLC, PDS, PDW, PDiaW, PDDiaW, PPN, PDPN	0.4554
NSen, NWo, NCha, ASLS, AWLC, PDS, PDW, PSVW, PDSNW, PPN, PDPN	0.4203
Baseline (NSen, NWo, NSyl, NDWo, AWLS, AWLC, PDS, PDW)	0.4501
NSen, NWo, NSyl, NDWo, AWLS, AWLC, PDS, PDW, PSVW, PDPN	**0.4770**

TABLE V
CLASSIFICATION RESULTS PERFORMED ON THE GROUPED-BY-2-CONTINUOUS-GRADES DOCUMENTS.

FEATURES	ACCURACY
Baseline (NSen, NSyl, NDSyl, AWLS)	0.6036
NSen, NSyl, NDSyl, AWLS, PSVW, PPN	0.6306
NSen, NSyl, NDSyl, AWLS, PDiaW, PDDiaW, PSVW, PDSVW	**0.6307**
Baseline (NSen, NSyl, NDSyl, PDS, PDW)	0.6173
NSen, NSyl, NDSyl, PDS, PDW, PSVW, PDSVW, PPN, PDPN	**0.6658**

TABLE VI
CLASSIFICATION RESULTS PERFORMED ON THE GROUPED-BY-SCHOOL DOCUMENTS.

FEATURES	ACCURACY
Baseline (NSen, NCha, NDWo, ASLW)	0.8246
NSen, NCha, NDWo, ASLW, PDiaW, PDDiaW, PPN, PDPN	0.8354
NSen, NCha, NDWo, ASLW, PDSVW, PDPN	**0.8434**
Baseline (NSen, NCha, AWLS, PDS, PDW)	0.8274
NSen, NCha, AWLS, PDS, PDW, PSVW, PDSVW	**0.8517**

country of Vietnam stretches over 3000 km with many different regions, each region has its own culture and language usage. Many regions have private words used only in that region but not in other places. Therefore, with the general text, especially the textbook, the appearance of the dialect words might affect the readability of the text. Similarly, the **Percentage of distinct dialect words (PDDiaW)** is also examined.

- **Percentage of Proper Noun (PPN):** In the text, the more the number of Proper Noun (person names, place names), the more the reader needs to memorize to identify exactly which person or place are mentioned. Therefore, the Percentage of Proper Noun in the text can also affect the readability of the text. In this study, we also examined the **Percentage of Distinct Proper Noun (PDPN)** for experiments.

IV. EXPERIMENTS

In this research, we used Support Vector Machines (SVM) to classify texts by readability: the Grade level or the Group-of-2-Grade level or the School level. In order to avoid over-fitting, we used k-fold (k=10) cross

validation for training and testing: the data sets are randomly divided into 10 parts: 9 parts are for the model training and the rest for testing. The models proposed by Luong et al. [14] are used as baselines. Table IV, V and VI show the result of our model in comparison with the baseline ones.

As can be seen in the Table IV, V and VI, when combine with our proposed features, most of the classification results are increased from 1–6% depending on the feature set compared to the baselines.

V. Conclusion

In this paper, we proposed a new method for classifying Literary Texts in Vietnamese Textbooks by Text Readability. The method is based on the use of Part-of-Speech features and specific features of Vietnamese language like the ratio of Sino-Vietnamese words, the ratio from the dialect words. The experiment results indicate that our method remarkably outperform the recently baselines of Luong et al. [14] and the features used are valuable.

For the future works, we will examine deeper features like syntactic, discourse, semantic, *etc*. to create more precise classifiers. Texts in other textbooks/domains will be collected and examined to build other specific models for Vietnamese text Readability assessment.

References

[1] J. D. Brown, G. Janssen, J. Trace, and L. Kozhevnikova, "A preliminary study of cloze procedure as a tool for estimating English readability for Russian students," in *Second Language Studies Paper*. University of Hawai'i at Manoa, 2012, pp. 1–22.

[2] E. Dale and J. S. Chall, "The concept of readability," *Elementary English*, vol. 26, no. 1, pp. 19–26, 1949.

[3] R. Flesch, *The Art of Readable Writing*. New York: Harper and Brothers Publishers, 1949.

[4] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Technical Training*, vol. Research B, no. February, p. 49, 1975.

[5] G. Robert, *The technique of clear writing*. New York: McGraw-Hill Book Co., 1952.

[6] H. M. Laughlin, "SMOG Grading-a New Readability Formula," *Journal of Reading*, vol. 12, no. 8, p. 639–646, 1969.

[7] L. Si and J. Callan, "A statistical model for scientific readability," in *Proceedings of the Tenth International Conference on Information and Knowledge Management*, ser. CIKM '01. New York, NY, USA: ACM, 2001, pp. 574–576.

[8] S. E. Schwarm and M. Ostendorf, "Reading level assessment using support vector machines and statistical language models," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 523–530. [Online]. Available: https://doi.org/10.3115/1219840.1219905

[9] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi, "Combining lexical and grammatical features to improve readability measures for first and second language texts," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York: Association for Computational Linguistics, April 2007, pp. 460–467. [Online]. Available: http://www.aclweb.org/anthology/N/N07/N07-1058

[10] K. Tanaka-Ishii, S. Tezuka, and H. Terada, "Sorting texts by readability," *Comput. Linguist.*, vol. 36, no. 2, pp. 203–227, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1162/coli.09-036-R2-08-050

[11] S. Vajjala and D. Meurers, "On improving the accuracy of readability classification using insights from second language acquisition," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 163–173.

[12] L. T. Nguyen and A. B. Henkin, "A Readability Formula for Vietnamese," *Journal of Reading*, vol. 26, no. 3, pp. 243–251, 1982. [Online]. Available: http://www.jstor.org/stable/40031716

[13] ——, "A Second Generation Readability Formula for Vietnamese," *Journal of Reading*, vol. 29, no. 3, pp. 219–225, 1985. [Online]. Available: http://www.jstor.org/stable/40029662

[14] A.-V. Luong, D. Nguyen, and D. Dinh, "Examining the text-length factor in evaluating the readability of literary texts in vietnamese textbooks," in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, Oct 2017, pp. 36–41.

[15] D. Dinh and D. D. Hao, "Chữ quốc ngữ hiện nay qua các con số thống kê (Current National Vietnamese language through statistics)," in *Hội thảo cấp Quốc gia về chữ quốc ngữ: sự hình thành, phát triển và những đóng góp vào văn hóa Việt Nam (National workshop about National Vietnamese language: the formation, development and contributions to Vietnam culture)*, PhuYen, Vietnam, Oct 2015.