Building a Corpus for Vietnamese Text Readability Assessment in The Literature Domain

An-Vinh Luong[1,3,*], **Diep Nguyen**[2,3], **Dien Dinh**[1,3]

[1]Computational Linguistics Center, University of Science, Ho Chi Minh City, Vietnam
[2]Department of Linguistics, University of Social Sciences & Humanities, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam

Received June 5, 2020; Revised August 4, 2020; Accepted August 10, 2020

Cite This Paper in the following Citation Styles
(a)*: [1] An-Vinh Luong, Diep Nguyen, Dien Dinh , "Building a Corpus for Vietnamese Text Readability Assessment in The Literature Domain," Universal Journal of Educational Research, Vol. 8, No. 10, pp. 100-109, 2020. DOI: 10.13189/ujer.2020.081073.*
(b)*: An-Vinh Luong, Diep Nguyen, Dien Dinh , (2020). Building a Corpus for Vietnamese Text Readability Assessment in The Literature Domain. Universal Journal of Educational Research, 8(10), 100-109. DOI: 10.13189/ujer.2020.081073.*

Abstract Text readability is a measure of how easy or difficult it is to read a text. This readability factor plays a crucial role in the processes of drafting and comprehending the texts, affecting the choice of proper texts for reading. Studies on the readability of text have started since the late nineteenth century and there have been many practical applications. However, these studies are mainly performed in English and other popular languages. In Vietnamese, the study of the text readability is still relatively untapped and has only received attention in recent years in the process of improving the curriculum and teaching methods. Recent studies on the readability of text in Vietnamese language are still limited, the main reason was largely due to the lack of text resources, which are corpora graded accordingly to difficulty levels. Therefore, in this study, we focused on building a corpus for assessing the readability of Vietnamese texts in the literature domain through the process of collecting, processing and evaluating documents. The result is that we have built up a corpus of 1,825 Vietnamese texts, divided into four levels of difficulty (Very easy, Easy, Medium and Difficult). Experiments with the existing Vietnamese readability assessment methods show that the built corpus is reliable and usable for further research on the text readability.

Keywords Text Readability, Vietnamese Language, Vietnamese Text Readability, Text Readability Corpus

1 Introduction

Reading is one of the fundamental skills for humans to acquire knowledge all over the world. Not only does reading give us new knowledge, but it also helps us relieve stress, develop intellectual capacities, help promote the brain, and even slow down the aging process [1]. However, in the modern days, billions of the texts, including both comprehensible and incomprehensible documents, are available for us to read. Therefore, finding out whether a text is suitable for a reader is of great significance in practice. There are two critical factors that determine the readability of a given text: the reader's ability and the text's difficulty itself. On the one hand, the reader's ability factor is related to the reader's prior knowledge (how much he knows about the subject before reading the text), his reading skill, his interest in the subject, and his motivation when reading the text. On the other hand, the text factor is related to its content (how the ideas are expressed, the style of expression, its organization, and navigational aids) and the way it is designed (layout, typography) [2].

Readability is one of the directions of the research of matching texts to readers. According to Brown et al. [3], text readability plays a crucial role in reading and understanding text, as it conveys the level of reading difficulty that a text may appear to native speakers. It can be expressed with an index, which is essentially a numeric scale approximating the degree of which a given text can be comprehended. Based on the readability, readers can determine whether a text is suitable for their reading ability. Besides, the author of a text can manipulates its readability to guide the reading flow and adjust the readability to the expected readership.

Many studies have only been conducted in English and other resources-rich languages like Arabic, Italian, French, Chinese, Japanese, among others. Al-Ajlan et al. [4] collected 60 Arabic texts from the 3rd to the 6th level of the elementary curricula and the 3rd level of the intermediate curriculum for evaluation. Chen et al. [5] examined a collection of 637 documents collected from textbooks of primary schools from grade 1 to grade 6 in the Chinese mainland. Dell'Orletta et al. [6] examined the corpus for readability features on both the text and the sentence levels. Their corpus was built from two sources: (1) a newspaper, La Republican; and (2) an easy-to-read newspaper, Due Parole. Chen and Daowadung [7] investigated 1,080 texts from 6 core subjects in the primary schools, Thailand, for their evaluations.

Nevertheless, the topic of text readability in Vietnamese, a resources-low language, remains relatively understudied, mainly due to the lack of necessary resources for examining and evaluating. In the first study in 1982, Nguyen and Henkin collected 20 passages from Vietnamese novels, magazines, and textbooks (from grade 4 to college), all of which were judged by 10 former Vietnamese teachers, teaching in the Vietnamese language. Then the authors examined these texts to develop the first formula for Vietnamese readability assessment [8]. In the second research, in 1985, with a similar approach, they collected 24 passages from Vietnamese textbooks and famous novels to create the second Vietnamese readability formula. Furthermore, 30 other passages were additionally used to evaluate this formula [9]. In the recent studies on Vietnamese text readability, Luong et al. [10, 12, 11], Diep et al. [13] examined around 380 texts collected from school textbooks to examine the effect of the text length and some specific Vietnamese language features on the text readability. In recent years, along with the process of the education reforms, the issue of assessing the readability of the text has received more attention. Therefore, it is an urgent need to build larger corpora for examining and assessing the readability of Vietnamese texts.

The aim of this study is describe the process of building a corpus for Vietnamese readability assessment in the field of literature. The article is organized as follows: Section 2 states the criteria for building the corpus; The process of building a corpus for Vietnamese readability assessment along with basic statistics and some experiments are presented in Section 3; Deeper statistics and analysis of the corpus are included in Section 4; Section 5 presents our experiments on the constructed corpus to check the reliability of the corpus; Section 6 concludes the study.

2 Criteria for building the corpus

Studies on text readability have two main directions [2]: (1) the printing elements direction, and (2) the content ones. The first direction focuses on printing characteristics of text like typeface, font color, font size, background color, graphics, charts, *etc*. Generally, researches on this direction use another term — legibility. The second one — content — mainly examines content elements like words usage, sentences usage, ideas organizing, *etc*. In this article, we built a corpus for Vietnamese readability assessment that only focuses on the content of the text without regard to the printing elements.

Several methods of building corpora for readability assessment have been proposed [14], each of which has its own strengths and weaknesses. The popular methods are follows:

- Expert judgments: Texts are evaluated by experts. The result of the text leveling is the texts, agreed by most or all the experts. This advantage of this method lies in its high reliability, but inviting the experts is carefully considered. This was also the method of Slyh & Hansen [15], Todirascu et al. [16], Xia et al. [17], Vajjala & Lučić [18], Nguyen & Henkin [8, 9], among others.

- Non-expert judgments: This is a method like the Expert judgments above, but texts are evaluated by a great number of regular readers instead of the experts. This method has proved the high reliability, but each of the texts must be evaluated by many the readers (at least 10 readers). This method was used by De Clercq et al. [19], Sinha & Basu [20], *etc*.

- Texts from textbooks: Texts are directly collected from the readings in textbooks and have already been evaluated in the book compiling process. Although the method has proven its efficiency for not requiring any experts, the number of collected texts is limited, and the copyright must also be taken into consideration. This method has been implemented in various works, including but not limited to Sun et al. [10], Dell'Orletta et al. [6], Chen & Daowadung [7], Lee & Hasebe [22], Berendes et al. [23], Luong et al. [10, 12], Diep et al. [13] *etc*.

- Comprehension test: People with different reading levels will be asked to read texts and provide their answers to questions related to the content of the texts, showing their understanding of the materials. The average comprehension level of the readers shows the difficulty of a given text. The advantage of this method is that it correctly reflects the real readability of a text by the evaluation of readers at different levels. However, giving correct and appropriate questions requires investigators that possess high reading skills and have clear insights regarding different levels of readers. This method was used by Yaneva et al. [24], Cutting & Scarborough [25], Keenan et al. [26] and Coleman et al. [27], among others.

- Cloze test: Like the comprehension test, but the readers are required to read and fill in some gaps, where the words were previously removed, to determine the ratio of gaps filled. This ratio is also used as the extent to which participants comprehend the text. This method has been widely employed as in the research of O'Toole & King [28], Gellert & Elbro [29], Trace et al. [30], among others, and can be applied to smaller forms of text, such as sentences and words. However, to ensure that the assessments are objective and accurate, this method does require many participants.

With the goal to build a large corpus effectively under our conditions, we integrated some methods mentioned above in this article. We mainly used (1) collecting texts and (2) expert evaluation to build the corpus.

The corpus was built based on the following criteria:

- Focused: The corpus must focus on one domain. In this work, we chose literature because these texts are easily collected on different levels of difficulty, and the readers do not need to have much specialized knowledge to be able to understand as in other fields. Therefore, the corpus should include: children's stories, short stories, novels, theoretical works on literature and language, *etc*. We did not collect poetry and drama texts because these texts contain either incomplete sentence structures (poetry) or various specialized structures (for scenes and roles in plays).
- Reliability: The corpus must be cross-evaluated by the experts to make sure that the texts were reliably collected.
- Large number: The total number of texts is large enough to make the corpus reliable during the experiment. This can also avoid over-fitting when using machine learning models.

3 Corpus building

The overall process is shown in Figure 1:

3.1 Collecting texts

In this study, we collected texts from some available sources, including but not limited to college-level textbooks, stories, and literature websites. Since most of the school textbooks are not digitally available, it took a lot of time and effort to process the texts. First, we collected them manually by scanning from the hard-copy versions. After that, we converted them into the digital format (using Optical Character Recognition — OCR), and then manually edited them. With other sources, they are digitally available on some websites, so we used the web crawling tool ParseHub to automatically collect them.

3.2 Pre-processing

After all the texts are collected, we pre-processed them. The process contained the following tasks: Text extraction: The source texts may contain both the text and non-text elements, such as images, tables, formulas, or even HTML tags (because many texts were crawled from the Internet). Therefore, we needed to remove these non-text elements. These non-text elements were mostly removed automatically by some computer scripts. However, since there are still some other elements that are not automatically removed, we did it manually.

- Spelling correction: Spelling errors are unavoidable in any corpora, especially corpora with a large amount of data. This corpus is not an exception. There are many errors needed to be corrected, especially the texts achieved by OCR. With these texts, we read all of them to correct their spellings manually without using any automatic spelling correction tool. As is well-known, most of the spelling correction tools are based on n-gram for detecting and correcting errors, and these tools only work well if spelling errors are not too close to each other. However, the texts obtained through OCR were full of errors in a sequence, and thus the automation tools were inappropriate in this case.
- Punctuation standardization: Punctuation like dot (.), comma (,), semi-colon (;), colon (:), exclamation (!), question (?), single quotation (‘), double quotation (“), brackets ([ ],(),), hyphen (-), slash (/), *etc*. were separated from their previous words by a space (“ ”). This made the texts clearer and enabled the statistical operations in these texts to be more exact.
- Encoding standardization: We standardized the data because the texts were collected from various sources with different encoding methods. For example, the Vietnamese word ’hoc’ (study) has 3 characters: ’h’, ’o’, ’c’ when this word is encoded in the pre-built Unicode. However, if it is done in the composite Unicode, this word includes four characters: ’h’, ’o’, ’c’, and ’.’ (drop-tone). In this work, we converted all of the documents into the pre-built Unicode because the ’tones’ in the Vietnamese language are also the elements affecting the text readability.
- Tone standardization: Similar to encoding, in Vietnamese, writing text have 2 types of placing the tone mark: the “old style” emphasizes aesthetics by placing the tone mark as close as possible to the center of the word (by placing the tone mark on the last vowel if an ending consonant part exists and on the next-to-last vowel if the ending consonant does not exist, as in ’hoa’, ’huy’); and the ’new style’ emphasizes linguistic principles and tries to apply the tone mark on the main vowel (as in ’hoa’, ’huy’). In this work, we converted all texts to the ’old style’.
- Sentence segmentation and word segmentation: Sentences and words are two common factors of readability research, often being examined in most readability studies — especially in readability formulas. They are also the basic factors for the other elements, such as part-of-speech (POS), named-entity (NE), dependency tree or lexical chain, *etc*. Consequently, the texts should be segmented sentences and segmented words.
- Document filtration: some documents still had too many errors (spelling, punctuation, encoding, *etc*.) after pre-processing (manually or automatically), so we proceeded to filter them out of the corpus.

In this research, we used the tool “CLC_VN_Toolkit” for punctuation, encoding, tone standardization, as well as sentence and word segmentation. This is a tool of COMPUTATIONAL LINGUISTICS CENTER[1] including the functions for pre-processing texts, such as sentence segmentation, word

[1] CLC — http://www.clc.hcmus.edu.vn (University of Science - VNU-HCM)

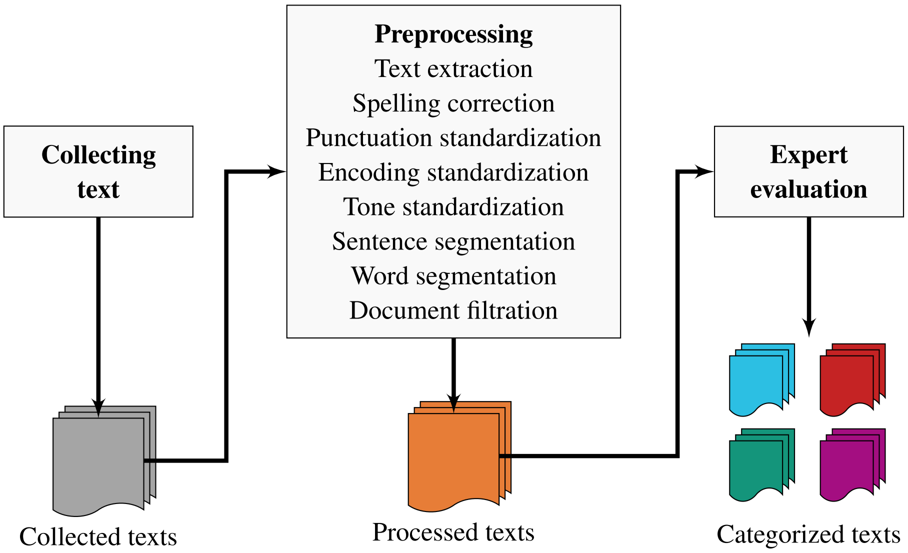

Figure 1. Corpus building process.

segmentation, part-of-speech tagging, named-entity recognition, *etc*. After all of the documents were pre-processed, the remaining corpus contained about 1,825 texts with the average length is 79.32 sentences.

3.3 Expert evaluation

In an attempt to make the corpus reliable, we invited some experts for evaluation. Twenty teachers who are teaching Vietnamese literature at middle and high schools, as well as colleges, were invited to evaluate the collected texts. The collected texts were sent to the experts to evaluate the text readability. Each text was ensured to be assessed by three experts. Each person was asked to read the given texts and categorizes each text into one of four categories as follows:

1. **Very easy:** Texts intended for children, and can be read and understood by those who have already or are currently in primary school
2. **Easy:** Texts intended for middle school students or people with middle school education, and can be read and understood by most of the people who have already or are currently in pre-intermediate education
3. **Medium:** Texts intended for high school students or people with high school education, and can be read and understood by most of the people who have already or are currently in intermediate education
4. **Difficult:** Specialized texts, especially ones that are intended for college students, as well as documents for specialized education, which can be read and understood by people who have already or are currently in advanced education.

If a text was categorized into different categories, that texts will be reviewed by a group of four experts, including five previous experts and two additional experts to decide the final category. The detailed statistics of the corpus are presented in Table 1.

To ensure the reliability of the evaluating process, we calculated the Fleiss' kappa score [31]. This is a statistical measure for assessing the agreement among a fixed number of the experts when classifying items. The Fleiss' kappa measure is usually used in natural language processing to assess the reliability of agreement among references in the corpus or among manual annotations by experts. The K value of our expert evaluation is 0.73, demonstrating a substantial agreement among the experts. As a result, the evaluating process is reliable, and the corpus is acceptable.

4 Corpus annotation and statistics

After all the collected texts were classified into different difficulty classes, we conducted some annotation steps and made some statistics on the corpus:

- Part-of-speech tagging: Some POS elements might have a good effect on text readability assessing models. Therefore, in this work, we conduct the POS tagging step for all classified text to use as features for further studies on the corpus. Table 2 presents the average number of each POS tag in the corpus (taking the sum of all the words that are labelled as that POS tag and divided it by the number of texts in the category) and the correlation coefficient between the ratio of that POS tag in the document and the readability level of that document. As shown below, the ratio of common nouns, volatile verbs, quality adjectives, prepositions and parallel conjunctions are highly correlated with the readability level of the documents, so these characteristics might be good for further studies on Vietnamese literature readability.
- Named-Entity tagging: NE elements also play an important role in the reading process. If a text has more NEs,

Table 1. Detailed statistic of the corpus

	Very easy	**Easy**	**Medium**	**Difficult**	**Overall**
No. texts	809	453	242	321	1,825
No. sentences	12,698	31,368	44,988	55,720	144,774
No. words	136,703	424,320	603,447	136,8142	253,2612
No. words and punctuations	809	453	242	321	1,825
No. syllables	155,406	489,668	703,530	1,874,914	3,223,518
No. texts	809	453	242	321	1,825
Avg. no. sentences	15.70	69.25	185.90	173.58	79.33
Avg. no. words	168.98	936.69	2,493.58	4,262.12	1,387.73
Avg. no. words and punctuations	197.33	1,082.00	2,889.39	4,964.50	1,612.40
Avg. no. syllables	192.10	1,080.94	2,907.15	5,840.85	1,766.31
Avg. no. syllables and punctuations	220.46	1,226.27	3,302.99	6,543.35	1,991.01
Avg. no. unique words	101.19	385.00	710.16	1,041.93	417.85
Avg. no. unique syllables	112	411	711	928	409

Table 2. Average number and the correlation coefficient of each POS tag in the corpus

Level	**Very easy**	**Easy**	**Medium**	**Difficult**	**Over-all**	**Correlation**
Proper Nouns	4.83	20.21	55.10	141.52	39.36	-0.56645
Countable Nouns	5.24	27.91	76.07	125.03	41.33	-0.68457
Concrete Nouns	0.67	4.09	9.48	13.55	4.95	-0.27278
Temporal Nouns	3.87	22.04	59.91	83.59	29.83	-0.60085
Numerals	7.57	39.99	105.35	242.37	69.88	0.76122
Common Nouns	45.81	242.04	618.70	1250.17	382.32	**0.86613**
Directional Verbs	0.83	4.65	10.60	5.12	3.83	-0.38593
State Verbs	2.09	11.86	30.41	48.99	16.52	0.59688
Comparative Verbs	1.55	8.69	25.48	82.13	20.67	0.70897
Volatile Verbs	34.70	195.03	499.74	728.64	258.22	**-0.81823**
Directional co-verb	2.03	9.78	25.35	9.64	8.39	-0.56541
Quantity Adjectives	0.22	1.48	3.89	9.94	2.73	0.34124
Quality Adjectives	13.24	71.04	194.14	351.78	111.12	**0.80346**
Demonstrative Pronouns	2.81	17.65	51.92	75.62	25.81	-0.61966
Personal Pronouns	9.60	53.76	154.57	113.43	58.05	-0.71007
Adverbs	14.57	90.30	247.44	259.82	107.39	-0.79893
Prepositions	7.26	41.67	118.62	365.57	93.59	**0.83726**
Parallel Conjunctions	9.06	56.83	155.56	300.58	91.62	**0.82781**
Subordinating Conjunctions	0.66	4.67	12.12	18.05	6.24	0.40682
Modifiers	1.71	9.75	29.71	25.13	11.54	-0.51015
Emotion Words	0.57	2.30	6.50	0.76	1.82	-0.33272
Foreign Words	0.03	0.26	1.38	9.29	1.90	0.23884
Onomatopoeia	0.01	0.06	0.26	0.06	0.07	-0.03679
Idioms	0.05	0.66	1.24	1.15	0.55	0.08534

Table 3. Average number of each NE tag in the corpus

Level	**Very easy**	**Easy**	**Average**	**Difficult**	**Over-all**
Person	4.01	15.40	42.25	81.14	25.48
Location	0.56	3.12	7.40	40.45	9.12
Organization	0.21	0.88	2.88	12.74	2.93
Brand	0.03	0.21	0.56	2.05	0.50
Number	3.51	18.17	45.76	84.33	26.96
Title	0.54	2.08	4.72	2.50	1.82
Date time	0.78	3.45	10.76	28.80	7.69
Measurement	0.33	2.02	5.19	4.21	2.08
Designation	0.11	0.54	1.07	4.62	1.14
Abbreviation	0.01	0.08	0.40	3.31	0.66
Terminology	0.02	0.02	0.03	0.16	0.04

Table 4. Common words and Vietnamese language characteristics statistics and the correlation coefficients

Level	Very easy	Easy	Medium	Difficult	Over-all	Correlation
No. common words	134	737	1948	2556	950	-
Ratio of common words	0.79	0.78	0.76	0.60	0.75	-0.92
No. unique common words	75	247	387	367	211	-
Ratio of unique common words	0.76	0.65	0.55	0.36	0.63	-0.93
No. common syllables	183	1038	2800	5617	1698	-
Ratio of common syllables	0.96	0.96	0.96	0.96	0.96	0.93
No. unique common syllables	106	384	647	814	371	-
Ratio of unique common syllables	0.95	0.94	0.92	0.88	0.93	0.91
No. dialect words	12.61	58.47	148.18	162.66	68.36	-
Ratio of dialect words	0.08	0.06	0.06	0.04	0.06	-0.74
No. unique dialect words	6.87	23.55	39.74	32.83	19.93	-
Ratio of unique dialect words	0.07	0.06	0.06	0.03	0.06	-0.76
No. Sino-Vietnamese words	35	203	556	1493	402	-
Ratio of Sino-Vietnamese words	0.21	0.22	0.24	0.35	0.24	0.86
No. unique Sino-Vietnamese words	23	105	219	468	147	-
Ratio of unique Sino-Vietnamese words	0.22	0.27	0.32	0.45	0.29	0.92
No. borrowed words	43	243	654	1609	449	-
Ratio of borrowed words	0.26	0.26	0.28	0.38	0.28	0.89
No. unique borrowed words	26.58	114.75	234.10	483.18	156.30	-
Ratio of unique borrowed words	0.26	0.30	0.34	0.47	0.32	0.92

then the reader would have to brainstorm more to remember and deduce which NEs are pointing to which specific entity. The detailed statistics are presented in Table 3. The results show that in difficult documents, the average number of NE is much more than the easy ones, for instance, in the Very easy texts, on average, only about 4 words are labelled as Person in each text, but there are an average of 81 words labelled as Person in each Difficult text. However, when conducting a correlation test, we found that the ratio of the number of each NE type divided by the number of words in the text does not have a strong correlation with the difficulty level of the text. This is because the harder the text is, the more words there are. Therefore, even though the number of each type of NE increases, the ratio of word count is virtually unchanged across each level.

- Statistics about common words: In many studies, the ratio of difficult words is an important feature when evaluating the text readability. However, creating the difficult word lists need a lot of effort, so most studies have used frequency lists as an alternative solution. That is, if a word does not appear in the frequency list, it will be considered a difficult word. In this study, we used the top 3,000 common words from the list of Dien et al. [33] as the frequency list. As indicated in the study of Dien et al., these 3,000 common words (10% of all Vietnamese words) occupy 90% of the word tokens appearing in Vietnamese texts. Consequently, these 3,000 words list can be considered to belong in the easy Vietnamese words list.

- Some Vietnamese language characteristics statistics: In this work, we also do other statistics about some specific Vietnamese language characteristics, such as:

  + The Vietnamese culture is strongly influenced by Chinese culture. The Vietnamese language is also affected, as more than 60% of Vietnamese vocabulary is derived from Chinese, known as Chinese-Vietnamese words. Sino-Vietnamese words are frequently used in scientific texts, technical texts, and formal texts, so Sino-Vietnamese words are often considered harder than other pure Vietnamese words. Therefore, the ratio of Sino-Vietnamese words was additionally used in this study. In this study, we proceed to extract features of Sino-Vietnamese words in the documents, using the list of Sino-Vietnamese words extracted from the Vietnamese Dictionary (2017) of Hoang Phe [34]. Words (including out-of-vocabulary words) which are not appear in this list are not treated as Sino-Vietnamese words.

  + The ratio of local words (dialect words): The country of Vietnam stretches over 3,000 km with many diverse regions, each of which has its own culture and language usage. Many regions retain private words habitually used 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. Similar to Sino-Vietnamese words, in this study, we also extracted dialect words from the Vietnamese Dictionary (2017) of Hoang Phe [34] for statistics. Words (including out-of-vocabulary words) which are not appear in this list are not treated as dialect words.

  + The ratio of borrowed words: Like Sino-Vietnamese words, many words used in Vietnamese originated from other languages such as English, French, Latin, *etc*., and they also often appear in formal texts. Therefore, we also include the statistics on the ratio of Vietnamese borrowed words. In this study, we also extracted the borrowed words from the Vietnamese Dictionary [34] as the basis for statistics.

The detailed statistics about common words and Vietnamese language characteristics are presented in Table 4. As shown below, the ratio of common words decreases gradually as the difficulty of the text gradually increases (0.79 in Very easy, 0.78 in Easy, 0.76 in Medium and 0.6 in Difficult texts). Conversely, dialect word rate, Sino-Vietnamese word ratio, and borrowed word gradually increase according to the difficulty level of the text. All the features are highly correlated with the difficulty level of the texts (the absolute value of the correlation coefficients is greater than 0.7).

5 Reliability testing

In order to re-examine the reliability of the corpus, we also performed the experiments. The first experiment is finding the correlation between the level of the text with the existing Vietnamese readability formulas.

The first formula was proposed in 1982 by Nguyen and Henkin [8]:

$$RL = 2WL + 0.2SL - 6 \tag{1}$$

where:

- RL: Readability Level of the text
- WL: Average Word Length in characters
- SL: Average Sentence Length in words.

The second formula for Vietnamese [9] was revised from the first one with the additional role of the ratio of difficult words in the document:

$$RL = 0.27WD + 0.13SL + 1.74 \tag{2}$$

where:

- RL: Readability Level of the text, the higher the RL value, the more difficult the text
- WD: Word Difficulty, achieved from the ratio of the number of compound Sino-Vietnamese words in the documents
- SL: Average Sentence Length in words.

Regarding the correlation, the difficulty of the text category correlates 0.76 with the results of the first formula and correlates 0.68 with the second formula. This means that the difficulty of the documents is highly correlated with the two formulas, and thus the corpus is reliable. Table 5 presents the readability score of the texts calculated by the two formulas. As shown below, the readability scores generally increase corresponding to the level of the categories. However, there are some documents that are a much higher or lower score than the group average. For example, the maximum score of a text in the Very easy category is 13.35 (calculated by the first formula), while the average score of that category is 6.89, which is even higher than the average score of the Difficult category. For further investigation, we examined that text and other similar texts — which were a much higher or lower score than the group average — and realized that these texts contained a lot of compound-words, which significantly increased the word length. The 13.35-scored-text only had 7 sentences but there were 202 words, including 76 compound words with many of the provinces and cities in Vietnam. That specific text was taken from grade 4 textbooks. Consequently, we noted that the formulas for measuring the readability of Vietnamese text are still limited and more research is needed to come up with methods to assess the readability of Vietnamese text more effectively.

In addition, we also used a machine learning method to evaluate the constructed corpus. This method is based on the study of Tanaka-Ishii et al. [32], which used Support Vector Machines (SVM) to create a model that compares and contrasts the readability of the text pairs based on word frequency features:

- First, all documents collected are extracted features for training models. Two types of word frequencies are used: Local frequency and Global frequency:
  - Local frequency: This is a basic statistical feature in many natural language processing problems. The local frequency of each word is calculated by the number of occurrences of that word divided by the total number of words in the text.
  - Global frequency: This is calculated by the log frequency of that word in a large corpus. We carried out the statistics on the Vietnamese Corpus (VCor) of Dien et al. [33], which includes 805K documents, 17M sentences, 346M words, and 18 topics/domains.

  Local and global frequencies are extracted for all words in the text and for all texts.

- Next, we constructed vectors of text pairs to train the comparison model. We create all 2-permutations of all the text: Each permutation is a pair of two texts (order matter) — for instance, a and b. We created a feature vector for each pair (V_{ab}) by taking the feature vector of text a (V_a) minus the feature vector of text b (V_b). Note that V_{ab} is not the same as V_{ba}, because if a is more difficult than b, the judgment for V_{ab} is 1, otherwise V_{ab} is -1. Therefore, we used permutation instead of combination. The result we had is $P^2_{|s|}$ feature vectors for training comparison models, where S is all the documents of the built corpus.

- Finally, we built a model for comparing the correlate readability of text pairs through the SVM classification algorithm. In this work, we used the K-Fold cross validation for the training and testing: splitting the vector set into 5 equal parts and use 4 parts for training and the rest for testing. The average accuracy rates of the comparison model are presented in Table 6. It can be seen that the accuracy is significantly high: over 97%, except for the results of the comparison between the texts in the Easy and the Medium categories (88%), and between the texts in the Easy and the Very easy categories, which is only 85%. Based on the survey, we found that despite being collected from the

Table 5. The readability score of the texts calculated by the two formulas

	The first formula			**The second formula**		
	Min	**Max**	**Average**	**Min**	**Max**	**Average**
Very easy	3.45	13.35	6.89	2.28	5.59	3.03
Easy	5.44	14.46	7.85	2.82	8.34	3.69
Medium	5.61	15.64	8.33	2.69	6.82	3.74
Difficult	8.13	17.59	13.02	3.45	8.34	5.11

Table 6. The accuracy rates of the comparison model (unit: %)

	Medium	Ease	Very easy
Difficult	99.76	99.95	99.50
Medium	-	88.35	97.21
Easy	-	-	85.15

more difficult category, some texts in the Medium category contained more common words when being compared to the other texts in the Easy and Very easy category. This led to some errors in the comparative results.

We also conducted an experiment on assessing the readability of the texts of the built corpus using the method of Luong et al. [12]: classify texts using the features of Number of sentences, Number of characters, Average word length calculated in syllables, Percentage of difficult words, Percentage of difficult syllables, Percentage of Sino-Vietnamese words and Percentage of unique Sino-Vietnamese words. The method used is also SVM and K-folds cross validation (we chose K=5). The accuracy of the experiment is 94.79%, proving that the set of documents we built has high accuracy and reliability.

6 Conclusions

In this article, we present the process of building a corpus for Vietnamese literature texts readability assessment. This is the first large corpus for evaluating readability for the Vietnamese language. We maximally utilized the online sources as well as the other sources with manual correction. To ensure that the corpus is reliable, we not only examined the corpus by statistics and quantitative methods, but also invited experienced experts for evaluation. The statistics show that our corpus is reliable and can be used for future studies about the readability assessment for Vietnamese texts. The corpus is available online for downloading at `https://github.com/anvinhluong/Vietnamese-text-readability`. We also conducted some experiments on this corpus, using the two existing Vietnamese text readability formula. Besides, we implemented a machine learning method to achieve a more effective result.

The results of this article confirm the need for more research on readability for the Vietnamese language. In the future, we will investigate this corpus with more common features along with some deeper Vietnamese characteristics in order to construct model(s) for the Vietnamese text readability assessment. Additionally, we will also build other domain-specific corpora for Vietnamese readability assessment to evaluate and compare the differences of the readability among the texts in various domains.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this article.

REFERENCES

[1] Wilson, R.S., et al., Life-span cognitive activity, neuropathologic burden, and cognitive aging. Neurology, 2013. 81(5): p. 314-21.

[2] Dubay, W.H., Smart Language: Readers, Readability, and the Grading of Text. 2007, Costa Mesa, California: Impact Information.

[3] Brown, J.D., et al., A preliminary study of cloze procedure as a tool for estimating English readability for Russian students, in Second Language Studies Paper. 2012, University of Hawai'i at Manoa. p. 1-22.

[4] Al-Ajlan, A.A., H.S. Al-Khalifa, and A.S. Al-Salman. Towards the development of an automatic readability measurements for arabic language. in 2008 Third International Conference on Digital Information Management. 2008. University of East London, London, UK: IEEE.

[5] Chen, Y.-T., Y.-H. Chen, and Y.-C. Cheng, Assessing Chinese Readability using Term Frequency and Lexical Chain. IJ-CLCLP, 2013. 18(2): p. 1-18.

[6] Dell'Orletta, F., et al. Assessing the Readability of Sentences: Which Corpora and Features? in Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications. 2014. Baltimore, Maryland: Association for Computational Linguistics.

[7] Chen, Y.-H. and P. Daowadung, Assessing readability of Thai text using support vector machines. Maejo International Journal of Science and Technology, 2015. 09(3): p. 355-369.

[8] Nguyen, L.T. and A.B. Henkin, A Readability Formula for Vietnamese. Journal of Reading, 1982. 26(3): p. 243-251.

[9] Nguyen, L.T. and A.B. Henkin, A Second Generation Readability Formula for Vietnamese. Journal of Reading, 1985. 29(3): p. 219-225.

[10] Luong, A.-V., 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). 2017.

[11] Luong, A.-V., D. Nguyen, and D. Dinh. A New Formula for Vietnamese Text Readability Assessment. in 2018 10th International Conference on Knowledge and Systems Engineering (KSE). 2018.

[12] Luong, A.-V., D. Nguyen, and D. Dinh. Assessing the Readability of Literary Texts in Vietnamese Textbooks. in 2018 5th NAFOSTED Conference on Information and Computer Science (NICS). 2018.

[13] Diep Thi Nhu, N., L. An-Vinh, and D. Dien, Affection of the part of speech elements in Vietnamese text readability. Acta Linguistica Asiatica, 2019. 9(1).

[14] François, T. An analysis of a French as a Foreign Language Corpus for Readability Assessment. in Proceedings of the third workshop on NLP for computer-assisted language learning. 2014. Uppsala, Sweden: LiU Electronic Press.

[15] Slyh, R. and E. Hansen, Detecting the Difficulty Level of Foreign Language Texts. 2010, Anticipate & Influence Behavior Division and Sensemaking & Organization Effectiveness Branch. p. 43.

[16] Todirascu, A., et al. Coherence and Cohesion for the Assessment of Text Readability. in Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2013). 2013. Marseille, France.

[17] Xia, M., E. Kochmar, and T. Briscoe. Text Readability Assessment for Second Language Learners. in Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. 2016. San Diego, CA: Association for Computational Linguistics.

[18] Vajjala, S. and I. Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. in Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. 2018. New Orleans, Louisiana: Association for Computational Linguistics.

[19] De Clercq, O., et al., Using the crowd for readability prediction. Natural Language Engineering, 2014. 20(3): p. 293-325.

[20] Sinha, M. and A. Basu, A study of readability of texts in Bangla through machine learning approaches. Education and Information Technologies, 2016. 21(5): p. 1071-1094.

[21] Sun, G., et al. Linear model incorporating feature ranking for Chinese documents readability. in The 9th International Symposium on Chinese Spoken Language Processing. 2014. Singapore: IEEE.

[22] Lee, J.H. and Y. Hasebe, Readability measurement for Japanese text based on leveled corpora. Papers on Japanese Language from an Empirical Perspective, Ljubljana: Academic Publishing Division of the Faculty of Arts, Univ. of Ljubljana, 2016.

[23] Berendes, K., et al., Reading Demands in Secondary School: Does the Linguistic Complexity of Textbooks Increase With Grade Level and the Academic Orientation of the School Track? Journal of Educational Psychology, 2017. 110(4): p. 518–543.

[24] Yaneva, V., et al. Combining Multiple Corpora for Readability Assessment for People with Cognitive Disabilities. in Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. 2017. Copenhagen, Denmark.

[25] Cutting, L.E. and H.S. Scarborough, Prediction of Reading Comprehension: Relative Contributions of Word Recognition, Language Proficiency, and Other Cognitive Skills Can Depend on How Comprehension Is Measured. Scientific Studies of Reading, 2006. 10(3): p. 277-299.

[26] Keenan, J.M., R.S. Betjemann, and R.K. Olson, Reading Comprehension Tests Vary in the Skills They Assess: Differential Dependence on Decoding and Oral Comprehension. Scientific Studies of Reading, 2008. 12(3): p. 281-300.

[27] Coleman, C., et al., Passageless Comprehension on the Nelson-Denny Reading Test: Well Above Chance for University Students. Journal of Learning Disabilities, 2009. 43(3): p. 244-249.

[28] O'Toole, J.M. and R.A.R. King, A Matter of Significance: Can Sampling Error Invalidate Cloze Estimates of Text Readability? Language Assessment Quarterly, 2010. 7(4): p. 303-316.

[29] Gellert, A.S. and C. Elbro, Cloze Tests May be Quick, But Are They Dirty? Development and Preliminary Validation of a Cloze Test of Reading Comprehension. Journal of Psychoeducational Assessment, 2012. 31(1): p. 16-28.

[30] Trace, J., et al., Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. Language Testing, 2015. 34(2): p. 151-174.

[31] Fleiss, J.L., Measuring nominal scale agreement among many raters. Psychological Bulletin, 1971. 76(5): p. 378-382.

[32] Tanaka-Ishii, K., S. Tezuka, and H. Terada, Sorting Texts by Readability. Comput. Linguist., 2010. 36(2): p. 203-227.

[33] Dinh, D., T.N. Nguyen, and H.T. Ho, Building a corpus-based frequency dictionary of Vietnamese. 2018. p. 72-98.

[34] Phe, H., Tu dien tieng Viet (Vietnamese dictionary). 8th ed. 2017: Da Nang Publishing House.