Assessing Vietnamese Text Readability using Multi-Level Linguistic Features

An-Vinh Luong[1], Dien Dinh[3]
Computational Linguistics Center
University of Science
Ho Chi Minh City, Vietnam

Diep Nguyen[2]
Department of Linguistics
University of Social Sciences & Humanities
Ho Chi Minh City, Vietnam

Thuy Bui[4]
Faculty of Foreign Languages
Hoa Sen University
Ho Chi Minh City, Vietnam

***Abstract*—Text readability is the problem of determining whether a text is suitable for a certain group of readers, and thus building a model to assess the readability of text yields great significance across the disciplines of science, publishing, and education. While text readability has attracted attention since the late nineteenth century for English and other popular languages, it remains relatively underexplored in Vietnamese. Previous studies on this topic in Vietnamese have only focused on the examination of shallow word-level features using surface statistics such as frequency and ratio. Hence, features at higher levels like sentence structure and meaning are still untapped. In this study, we propose the most comprehensive analysis of Vietnamese text readability to date, targeting features at all linguistic levels, ranging from the lexical and phrasal elements to syntactic and semantic factors. This work pioneers the investigation on the effects of multi-level linguistic features on text readability in the Vietnamese language.**

Keywords—Text readability; text difficulty; readability formula; linguistics features; Vietnamese

I. Introduction

Text readability is a measure of how easy or difficult a text is to be read [1], effectively guiding the process of comprehending that text. The readability of a document heavily depends on its linguistic features such as word usage, phrasal structures, and sentence meaning. Not only does text readability help readers determine whether a document is suitable to read, but it also assists authors in adjusting their writing for the target audience.

Building a model to assess the readability of texts yields great significance across various disciplines. In academia, researchers can rely on text readability to improve their scientific communications, while curriculum designers can be assured in developing appropriate course outlines for each age group of students, and language teachers can effectively create or select relevant second language learning materials for foreigners. Moreover, text readability plays a key role in aiding publishers in establishing varied audiences, supporting policy makers in drafting legal documents that accommodate all citizens with different literacy levels, and supporting manufacturers in preparing product manuals.

Research on the readability of text has been conducted since the late nineteenth century, with a special focus on English and other resource-rich languages. These studies are generally divided into two main approaches: the statistical approach and the machine learning approach. The statistics-oriented works mainly examine how the features of a text affect that text's readability using correlation and regression analyses. These analyses determine features that are highly correlated with readability and calculate the weight of those features, respectively, to develop formulas that predict the readability of that text. Representative works of this approach include the Dale-Chall formula [2], the SMOG formula [3], among others. Meanwhile, studies that follow the machine learning approach seek to exploit neural network algorithms with great computational power that enable the manipulation of a broader range of features and at a deeper level to create text classifiers based on the readability level. Works that demonstrate this approach are Si and Callan [4], Collins-Thompson and Callan [5], Pitler and Nenkova [6], Vajjala and Meurers [7], Sinha and Basu [8], Vajjala and Lučić [9], and Al Khalil, et al. [10], among others.

In Vietnamese, research on text readability remains relatively limited. First, Nguyen and Henkin [11] pioneered this vein of research for overseas Vietnamese people. Then, in 2017, when examining the features of text in linguistic textbooks, Luong, et al. [12] showed that the text length significantly influences the classification of these grammatical texts by readability level. In another study in 2018, Luong, et al. [13] further argued that Sino-Vietnamese elements and dialect features also plays a critical role in evaluating the readability of texts in Vietnamese textbooks.

Besides the relatively small number of studies on this topic in Vietnamese, the features examined are only at shallow levels, with surface statistics such as word frequency and type-token ratio. Features at higher levels like syntax and semantics remain still untapped, mainly due to the lack of survey resources and the low accuracy rates yielded by in-depth word processing tools. Recently, more extensive studies on Vietnamese texts have gained increasing attention and promising results, leading to their application to the problems of natural language processing in general and the question of text readability in particular. Therefore, in this study, we investigate the effects of linguistic features on the readability of text in Vietnamese. These linguistic features range from word-level (word frequency, language, sentence length, etc.) and Language model features (bi-gram, tri-gram, etc.) to syntactic (parsing tree height/width, number of clauses, etc.)

and fundamental semantic features (average of semantic numbers of words/sentences). Not only is this work the most comprehensive study on this topic in Vietnamese as of the time of publication, it is also the first to exploit the deepest linguistic level of Vietnamese texts for the readability question.

The rest of the paper will be structured as follows: Section 2 presents relevant previous works addressing the text readability problem. Section 3 introduces the features examined, the dataset used for the examinations, the methods, and the results of our study. Finally, Section 4 contains the bulk of discussions and conclusions drawn from the experimental process.

II. Related works

In this section, we will introduce previous studies on the text readability problem in the world as well as in Vietnamese. As introduced in Section 1, the study of text readability has begun since the end of the nineteenth century. While a great deal of works has been published since then, the research focus has been on English and other resource-rich languages.

There are two main approaches in the study of text readability: (1) statistical approach and (2) machine learning approach. In the statistical approach, researchers focus mainly on identifying features closely related to the difficulty of a text through correlation analysis. Then, the selected features are used to construct the readability measurement formulas of the text. This approach has been implemented in a broad range of studies, including but not limited to Chall and Dale [2], Kincaid, et al. [14], Zeno, et al. [15], as well as Lee and Hasebe [16]. In Vietnamese, there have been four studies based on this approach: three of which are by Nguyen and Henkin [17], Nguyen and Henkin [11], Luong, et al. [18], and one of which is by Nguyễn, et al. [19].

TABLE I. Some notable studies in recent years on text readability for resource-rich languages and for Vietnamese

Work	Dataset	Features
Statistical approach		
Kincaid, et al. [14]	531 subjects from four schools at two Navy bases	Average length of sentences and average number of syllables per word
Chall and Dale [2]		Percentage of difficult words and average length of sentences
Lee and Hasebe [16]	A combination of texts from 83 introductory to advanced Japanese textbooks and texts from National Diet meeting transcripts, categorized into 6 scale levels	Average length of sentences, proportion of kango, proportion of wago, proportion of verbs, and proportion of auxiliary verbs
Machine learning approach		
Sun, et al. [23]	637 documents extracted from textbooks for grades one to six in mainland China	76 text features from surface features, Part-of-Speech features, parse tree features, and Entropy features
De Clercq, et al. [21]	105 paragraphs from the Dutch LassyKlein corpus	Fundamental level, language model features, and deeper level features
Chen and Daowadung [24]	720 texts from six subjects of elementary school textbooks in Thailand	Term frequency features, shallow features, and language model features
Berendes, et al. [22]	2,928 readings in the geographic textbooks of four publishers in Germany from grades 5 to 10	Vocabulary, syntax, morphology, and cohesion-related features
Tseng, et al. [25]	1,441 social science articles and 772 natural science articles	LSA features
Vietnamese		
Statistical approach		
Nguyen and Henkin [17]	20 text paragraphs with about 300 words each from Vietnamese novels and magazines, as well as textbooks of Vietnamese students in the United States from grade 4 to college	Average length of sentences and average length of words
Nguyen and Henkin [11]	24 text paragraphs with about 300 words each from Vietnamese novels and magazines, as well as textbooks of Vietnamese students in the United States from grade 4 to college	Word difficulty and average length of sentences
Luong, et al. [18]	996 texts collected from stories for children, sample essays, fairytales, textbooks, newspapers, political theory articles, language and literary articles, law, and legal documents,…	Average length of sentences, average length of words, and percentage of difficult words
Nguyễn, et al. [19]	209 prose texts in Vietnamese textbooks for elementary school children from grades 2 to 5	25 Part-of-Speech elements
Machine learning approach		
Luong, et al. [12]	288 texts from Vietnamese textbooks for elementary students and Literature textbooks for junior high school students in Vietnam	Average length of sentences, average length of words, and percentage of difficult words, and the length of text
Luong, et al. [13]	372 texts from Vietnamese textbooks for general students in Vietnam	Percentage of Sino-Vietnamese words, percentage of dialect words, and percentage of proper nouns

Meanwhile, in the machine learning approach, features are included in machine learning classifiers to evaluate which features help increase the accuracy of the classification process. Some pre-graded reference texts are utilized to train the model and evaluate the classification accuracy. Some of the notable studies on this approach are Dell'Orletta, et al. [20], De Clercq, et al. [21], and Berendes, et al. [22], etc. In Vietnamese, studies based on this approach have only been carried out in recent years like those of Luong, et al. [12], Luong, et al. [13].

Table I presents a summary of some influential studies on text readability from both approaches in recent years along with information about the dataset and features examined for a range of languages, including Vietnamese.

III. Research Design and methodology

In this section, we will present our examinations on linguistic features of documents that can be extracted automatically by word processing tools for Vietnamese (up to the present time) to address the question of assessing the readability of Vietnamese writings.

A. *Features*

In this study, we examined 271 linguistic features listed in Table II. These features range from superficial features such as the average sentence length, the ratio of Sino-Vietnamese words, and the local word ratio, etc. (21 features in total) and word-type (Part-of-speech – POS) level features, such as the ratio of proper nouns, the average number of word-types, etc. (150 features in total) to syntax-level features such as the depth of syntactic trees, the numbers of clauses and of connected words per sentence, etc. (31 features in total) and basic semantic features such as the ratios of monosemous words and of polysyllabic words, the average number of meaningful units per sentence, etc. (10 features in total). Regarding features at the shallow level, we examined 30 language model features such as the average rank, the average frequency, and the average perplexity value of n-grams. These n-grams include character n-gram, syllable n-gram, word n-gram at bi- and tri-grams levels. Meanwhile, for features at the word-type level, we focus on the language model features at word bi-grams and word tri-grams (12 features in total). At the semantic level, given that research on automatic semantic labeling in Vietnamese text is still limited, we only extracted 17 basic statistical features such as the ratios of monosemous words and of polysemous words, the average meaningful units per word in the text, as well as the geometric mean of meaning of sentences in text, etc.

TABLE II. List of features examined

RAW FEATURES		
distinct easy syllables/distinct syllables	ratio of monosyllabic words	
distinct easy word/distinct words	ratio of polyphonic words	average word length in character
ratio of 2-syllable words	average sentence length in character	average word length in syllable
ratio of 3-syllable words	average sentence length in syllable	ratio of long sentence (in syllable)
ratio of distinct easy syllables	average sentence length in word	ratio of long sentence (in word)
ratio of distinct easy words	average sentence lengths in syllable (remove	ratio of short sentence (in syllable)
ratio of easy syllables	duplicate)	ratio of short sentence (in word)
ratio of easy words	average sentence lengths in word (remove duplicate)	
PART-OF-SPEECH FEATURES		
POS tags/sentences	distinct directional verbs/distinct words	emotion words/distinct words
POS tags/words	distinct directional verbs/sentences	emotion words/sentences
ratio of 2-POS tag words	distinct directional verbs/words	emotion words/words
ratio of 3-POS tag words	distinct emotion words/distinct words	foreign words/distinct words
ratio of multi POS tag words	distinct emotion words/sentences	foreign words/sentences
ratio of single POS tag words	distinct emotion words/words	foreign words/words
adverbs/distinct words	distinct foreign words/distinct words	idioms/distinct words
adverbs/sentences	distinct foreign words/sentences	idioms/sentences
adverbs/words	distinct foreign words/words	idioms/words
common nouns/distinct words	distinct idioms/distinct words	modifiers/distinct words
common nouns/sentences	distinct idioms/sentences	modifiers/sentences
common nouns/words	distinct idioms/words	modifiers/words
comparative verbs/distinct words	distinct modifiers/distinct words	numerals/distinct words
comparative verbs/sentences	distinct modifiers/sentences	numerals/sentences
comparative verbs/words	distinct modifiers/words	numerals/words
concrete nouns/distinct words	distinct numerals/distinct words	onomatopoeia/distinct words
concrete nouns/sentences	distinct numerals/sentences	onomatopoeia/sentences
concrete nouns/words	distinct numerals/words	onomatopoeia/words
countable nouns/distinct words	distinct onomatopoeia/distinct words	parallel conjunctions/distinct words
countable nouns/sentences	distinct onomatopoeia/sentences	parallel conjunctions/sentences

countable nouns/words	distinct onomatopoeia/words	parallel conjunctions/words
demonstrative pronouns/distinct words	distinct parallel conjunctions/distinct words	personal pronouns/distinct words
demonstrative pronouns/sentences	distinct parallel conjunctions/sentences	personal pronouns/sentences
demonstrative pronouns/words	distinct parallel conjunctions/words	personal pronouns/words
directional co-verb/distinct words	distinct personal pronouns/distinct words	prepositions/distinct words
directional co-verb/sentences	distinct personal pronouns/sentences	prepositions/sentences
directional co-verb/words	distinct personal pronouns/words	prepositions/words
directional verbs/distinct words	distinct prepositions/distinct words	proper nouns/distinct words
directional verbs/sentences	distinct prepositions/sentences	proper nouns/sentences
directional verbs/words	distinct prepositions/words	proper nouns/words
distinct adverbs/distinct words	distinct proper nouns/distinct words	quality adjectives/distinct words
distinct adverbs/sentences	distinct proper nouns/sentences	quality adjectives/sentences
distinct adverbs/words	distinct proper nouns/words	quality adjectives/words
distinct common nouns/distinct words	distinct quality adjectives/distinct words	quantity adjectives/distinct words
distinct common nouns/sentences	distinct quality adjectives/sentences	quantity adjectives/sentences
distinct common nouns/words	distinct quality adjectives/words	quantity adjectives/words
distinct comparative verbs/distinct words	distinct quantity adjectives/distinct words	state verbs/distinct words
distinct comparative verbs/sentences	distinct quantity adjectives/sentences	state verbs/sentences
distinct comparative verbs/words	distinct quantity adjectives/words	state verbs/words
distinct concrete nouns/distinct words	distinct state verbs/distinct words	subordinating conjunctions/distinct words
distinct concrete nouns/sentences	distinct state verbs/sentences	subordinating conjunctions/sentences
distinct concrete nouns/words	distinct state verbs/words	subordinating conjunctions/words
distinct countable nouns/distinct words	distinct subordinating conjunctions/distinct-words	temporal nouns/distinct words
distinct countable nouns/sentences	distinct subordinating conjunctions/sentences	temporal nouns/sentences
distinct countable nouns/words	distinct subordinating conjunctions/words	temporal nouns/words
distinct demonstrative pronouns/distinct words	distinct temporal nouns/distinct words	volatile verbs/distinct words
distinct demonstrative pronouns/sentences	distinct temporal nouns/sentences	volatile verbs/sentences
distinct demonstrative pronouns/words	distinct temporal nouns/words	volatile verbs/words
distinct directional co-verb/distinct words	distinct volatile verbs/distinct words	
distinct directional co-verb/sentences	distinct volatile verbs/sentences	
distinct directional co-verb/words	distinct volatile verbs/words	
SYNTAX-LEVEL FEATURES		
average height of clauses (parse tree)	average no. clauses	average number of nonterminal nodes (parse tree)
average height of level 1 branches (parse tree)	average no. clauses (remove duplicate)	average number of noun phrases
average highest clauses (parse tree)	average no. conjunction word	average number of prepositional phrases
average length of clauses	average no. content words	average number of terminal nodes (parse tree)
average longest clauses	average no. distinct conjunction word	average number of verb phrase
average longest noun phrases	average no. function words	average tree breadths (parse tree - remove duplicate)
average longest preposition phrases	average no. level 1 branches (parse tree)	average tree breadths (parse tree)
average longest verb phrases	average no. level 1 nonterminal nodes (parse tree)	average tree depths (parse tree - remove duplicate)
average no. brackets (parse tree)	average no. nodes (parse tree - remove duplicate)	average tree depths (parse tree)
average no. branches (parse tree - remove duplicate)	average no. nodes (parse tree)	ratio of simple sentences
average no. branches (parse tree)		
BASIC SEMANTIC FEATURES		
ratio of 2-semantic words	average of word semantic/sentences	product of word semantics/words
ratio of 3-semantic words	geometric mean of word semantic/sentences	semantics/sentences
ratio of monosemous words	product of word semantics/sentences	semantics/words
ratio of polysemous words		
RAW-LEVEL LANGUAGE MODEL FEATURES		
average character bigram frequencies	average syllable bigram frequencies	average word bigram frequencies
average character bigram perplexity	average syllable bigram perplexity	average word bigram perplexity
average character bigram rankings	average syllable bigram rankings	average word bigram rankings
average character trigram frequencies	average syllable list frequencies	average word list frequencies
average character trigram perplexity	average syllable rankings	average word rankings
average character trigram rankings	average syllable set frequencies	average word set frequencies

average distinct syllable frequency average distinct word frequency average frequency of sentence length in syllable (remove duplicate) average frequency of sentence length in word (remove duplicate)	average syllable set rankings average syllable trigram frequencies average syllable trigram perplexity average syllable trigram rankings	average word set rankings average word trigram frequencies average word trigram perplexity average word trigram rankings
POS-LEVEL LANGUAGE MODEL FEATURES		
average POS bigram frequencies average POS bigram perplexity average POS bigram rankings average POS trigram frequencies	average POS trigram perplexity average POS trigram rankings average word with POS bigram frequencies average word with POS bigram perplexity	average word with POS bigram rankings average word with POS trigram frequencies average word with POS trigram perplexity average word with POS trigram rankings
VIETNAMESE-SPECIFIC FEATURES		
distinct borrowed words/distinct words distinct local words/distinct words distinct Sino-Vietnamese words/distinct words ratio of borrowed words ratio of distinct borrowed words ratio of local words ratio of distinct local words	ratio of Sino-Vietnamese words ratio of distinct Sino-Vietnamese words 2-syllable Sino-Vietnamese words/Sino-Vietnamese words 2-syllable Sino-Vietnamese words/words 3-syllable Sino-Vietnamese words/Sino-Vietnamese words	3-syllable Sino-Vietnamese words/words monosyllabic Sino-Vietnamese-words/Sino-Vietnamese words monosyllabic Sino-Vietnamese-words/words polyphonic Sino-Vietnamese words/Sino-Vietnamese words polyphonic Sino-Vietnamese-words/words

B. Corpus

Following most of the previous studies on text readability in Vietnamese, this study also used the corpus of 371 literature texts by Luong, et al. [13]. Moreover, the collection and construction of a new dataset for the survey are extremely costly in terms of time and labor, and thus utilizing this existing corpus the optimal option. The research on texts of other domains will be carried out in future studies.

These documents were collected from Vietnamese and Literature textbooks for students in Vietnam. All of these textbooks are written in Vietnamese and published by Vietnam Education Publishing House under the resolution to renovating the program for general education of the National Assembly, The Socialist Republic of Vietnam, in 2000 [26].

In Vietnam, primary education is divided into five years – from grade 1 to grade 5. However, the Vietnamese textbooks for first grade students only include reading and writing exercises for simple characters and words, and thus they were not included in the surveys. The textbook for junior high school students is categorized into four levels, corresponding to four school years – from grade 6 to grade 9. For high school students, the Literature textbooks are partitioned into three levels corresponding to three school years – from grade 10 to grade 12. The Literature textbooks for high school students are also classified into two different sets: (i) a general set for most students and (ii) an advanced set, with more reading, for students specialized in Literature. Table III presents the statistics of the corpus.

To extract the features that we mentioned in Section 3.1 for each document, we took steps to process and label the text. This process consists of the following steps:

Encoding standardization: We standardized the data because the texts were collected from various sources with different encoding methods. For instance, the Vietnamese word "học" (study) consists of three characters – "h", "ọ", "c" – when this word is encoded in the pre-built Unicode. However, if it is done in the composite Unicode, this word includes 4 characters: "h", "o", "c", and "." (drop-tone). In this article, we converted all the documents into the pre-built Unicode.

TABLE III. CORPUS STATISTICS

	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12
Number of documents	67	62	40	40	28	13	17	21	15	19	49
Average number of sentences	18.3	19.6	21.5	21.4	54.8	46.4	65.8	107.3	60.7	105.2	111.7
Average number of words	158	192	231	244	680	677	969	1447	862	1360	1710
Average number of syllables	178	222	276	288	784	821	1131	1710	1006	1579	2179
Average number of characters	827	1065	1335	1396	3709	3942	5402	8160	4860	7535	10761
Average number of distinct words	100.6	125.6	144.3	152.8	304.9	329.7	394.3	526.3	368.4	510	576
Average number 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

Punctuation standardization: Punctuation like the 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 enable the texts to appear clearer and the statistical operations in these texts to be more exact.

1) Tone standardization: Similar to encoding, in Vietnamese, there are two ways to place the tone mark. First, 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 “hóa”, “hủy”). Meanwhile, the “new style” emphasizes linguistic principles and applies the tone mark on the main vowel (as in “hoá”, “huỷ”). In this work, we converted all texts to the “old style”.

2) Sentence segmentation and word segmentation: Sentences and words are two common features of readability research, often being examined in most readability studies – especially in readability formulas. They are also the basic features for other elements, such as part-of-speech (POS), named-entity (NE), dependency tree, or lexical chain, etc. Consequently, the texts were segmented into sentences, which, in turn, were segmented into words.

3) POS tagging: POS features are commonly used in text readability studies, such as Vogel and Washburne [27], Bormuth [28], Al Khalil, et al. [10], among others. Therefore, in this study, we conducted the POS tagging for documents in preparation for extracting features in Section 3.1.

4) Constituency parsing: Syntactic features have been widely exploited in the literature on text readability in the world. However, for Vietnamese, due to limitations on syntax labeling tools and methods, the syntax features remain relatively unexplored, not only with the readability of the text, but also with various other problems in the field of the Vietnamese language. However, recently, the accuracy of studies on automatic constituency parsing in Vietnamese has been significantly improved. In particular, Uyen, et al. [29] has achieved an accuracy rate of 79%. In this study, to effectively examine the syntactic features that affect the level of text readability, we used the results of Phan et al.’s research to parse documents in the corpus.

In this study, we used the CLC_VN_TOOLKIT of the Computational Linguistics Center (CLC)[1] to preprocess, split sentences, separate words, and tag POS. The tool’s accuracy data was not disclosed, but our experiments indicates that the accuracy achieved was over 99% for the sentence and word tokenization tasks and over 97% for the POS tagging task.

After all the documents were processed and the necessary labels were assigned, we proceeded to extract the features for the examinations. The extraction of most of the features mentioned in Section 3.1 could be achieved straightforwardly from the processing and labeling steps. However, there were some features require additional support of external corpora, as follows:

1) Easy words and syllables features: In various studies, the ratio of easy words in a text remains a crucially dominant feature in the evaluation of the readability of that text. However, constructing a list of easy words is remarkably costly, as it requires a large number of readers to examine a large number of words. Hence, most studies commonly utilize frequency word lists instead. That is, if a word has a high frequency of use, it is likely that native speakers perceive that word as easy to understand, and vice versa. Likewise, easy syllable features were also implemented in this study. Our target is the readings in Vietnamese and Literature textbooks for students in Vietnam, and thus we used the list of the 3,000 most common words and the 3,000 most common syllables in Vietnamese of Dinh, et al. [30]. If a word appeared in this list of 3,000 common words, it would be treated as an easy word. Other words (including out-of-vocabulary (OOV) words) were treated as not-easy words. Similarly, a syllable was considered an easy syllable if it appeared in this list of 3,000 common syllables. It is possible for a word or syllable to appear more often only in a specific domain of text, and hence, are easier to comprehend to only a particular group of readers, but not to other text domains or other reader groups. In those cases, the list of frequent/easy words/syllables should be different.

2) Sino-Vietnamese features: The Vietnamese culture is strongly influenced by the 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, and they are often considered more difficult than other pure Vietnamese words. Therefore, the ratio of Sino-Vietnamese words was additionally used in this study. We extracted features of Sino-Vietnamese words in the documents using the list of Sino-Vietnamese words from the Vietnamese Dictionary by Phe [31]. Words (including OOV words) that did not appear in this list were not treated as Sino-Vietnamese words.

3) Dialect features: The country of Vietnam stretches over 3,000 km with various 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 general texts, especially textbooks, 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 by Phe [31] for statistics. Words (including OOV words) that did not appear in this list were not treated as dialect words.

4) Language model features: Language models are often implemented in a broad range of studies on NLP in general and on text readability in particular. Simply stated, a language model is a probability distribution over text sets, indicating how likely a sentence or phrase occurs in a language. The

[1] http://www.clc.hcmus.edu.vn

higher the probability of a sentence or phrase is, the more familiar that sentence or phrase is to the readers. Consequently, that sentence or phrase may be easier to read than the low probability sentence or phrase. In this study, to extract features for the text difficulty problem, we built several language models, which include characters, syllables, words, words with POS, POS-only bi-grams, and tri-grams. The corpus that we utilized to construct the language model is VCor (Vietnamese Corpus) [30]. This corpus consists of 805,000 documents, extracted from a broad range of sources such as news sites, books, and Vietnamese newspapers, etc.

5) Semantic features: Since there is no semantic corpus with sufficiently large quantity to conduct the examination and experiment, no previous studies have focused on the processing or automatic semantic labeling of sentences in Vietnamese. In this study, we extracted basic statistical semantic features, such as the average number of meanings of words in a sentence and Geometric Mean of meanings of sentences in a text, among others. We also used the Vietnamese Dictionary by Phe [31] to conduct statistics on the meaning of words and extract the features that we mentioned in Section 3.1.

6) Text grouping: In this study, we grouped documents in two ways to fit each approach of the text readability assessment problem and match our examination method:

a) By school track: Texts were grouped into three school tracks, which were elementary, middle, and high schools. We grouped documents in this way to conduct features examinations according to the feature evaluation method of the text classification problem.

b) By grade level: Texts were grouped into 11 grade levels according to the curriculum of the general textbook in Vietnam. With this grouping, we investigated the role of the features using correlation and regression analyses.

C. Features Examination

In this study, we conducted surveys that evaluate the impact of NLP features introduced in Section 3.1 on text readability. These evaluations were based on the examinations on the textbook materials for Vietnamese students mentioned in Section 3.2.

We implemented two examination methods corresponding to two approaches of the text readability assessment problem:

1) Statistical approach

This approach mainly implements correlation analysis to identify the features highly correlated with the readability level, thereby extracting the weight of these features through regression analysis method to build formula(s) to predict the difficulty of the texts. This was also the approach used for developing famous text readability formulas such as Dale-Chall [2], Flesch Reading Ease [14], SMOG [3], as well as the first and second formulas for Vietnamese text in Nguyen, et al. [11, 17].

Correlation analysis determines the linear relationship between the quantitative variables in this study, which are the features of the text and the readability level of that text. The higher the correlation coefficient between the two variables is, the higher the degree of their correlation is. The correlation coefficient ranges from -1 to 1. A correlation coefficient of 0 (or nearly 0) indicates that the two variables almost have no contact with each other. Conversely, a coefficient of -1 or 1 signals that the two variables have an absolute relationship. If the value of the correlation coefficient is negative ($r < 0$), it suggests that when the value of one variable increases, the value of the other decreases (and vice versa, when one variable decreases, the other increases). Meanwhile, if the correlation coefficient value is positive ($r > 0$), it means that when one variable increases, the other increases, and vice versa. In this study, we use the Pearson correlation coefficient. Table IV presents a list of features that are highly correlated with the readability level of the text (with a correlation coefficient greater than or equal to 0.8 or less than or equal to -0.8). These features consisted of 13 raw text features, 7 POS features, 2 syntax-level features, 3 basic semantic features, 21 raw-level language model features, 6 POS-level language model features, and 4 Vietnamese specific features. The raw-level language model features and 15 raw text features were most strongly correlated with the readability level of Vietnamese texts with the highest correlation coefficients being 0.91 and 0.85, respectively. Other features like POS, syntax, basic semantic, or Vietnamese specific features were not as strongly correlated as raw-level language model and raw features, but also had high correlation coefficients, from 0.80 to 0.84.

After correlation analysis, we selected features closely related to the difficulty of the text to perform regression analysis. Regression analysis is a statistical technique used to estimate the equation that best fits the set of observations of the dependent variable, which is the text readability level in this study, and the independent variables, which are the features used. Regression analysis allows the best estimation of the true relationship between variables. From this estimating equation, we can predict the dependent variable (the readability level of the text – unknown) based on the given value of the independent variable (the features – known). In regression analysis, if independent variables strongly correlated with each other (high correlation coefficient), multi-collinearity phenomenon will occur. Therefore, independent variables that are strongly correlated with each other are typically removed before the regression analysis. However, during the process of correlation analysis, we found that all the features in Table IV were strongly correlated with each other (the correlation coefficients were $\geqslant$ 0.7), and thus we conducted two experiments: (1) regression analysis with features in Table IV, with no exclusion of any strongly correlated features, and (2) regression analysis with features that correlate with the readability of text greater than or equal to 0.7, eliminating features that were strongly correlated with each other. We did not remove the strongly correlated features in the first experiment because the feature that had the highest correlation with the text readability level – average word set rankings – was also strongly correlated with the remaining features, with correlation coefficient values $\geqslant$ 0.8. For the second experiment, we selected the features with the correlation coefficient with text difficulty $\geqslant$ 0.7 and removed the features that correlated with the selected features ⩾ 0.8. As a result, the remaining number of features is only three. If we were to lower the elimination threshold to 0.7, only one feature with the highest correlation coefficient would have been chosen. Table V and Table VI present the intercept scores the coefficients of the features in the estimation equation after regression analysis of both experiments. Table VII shows the correlation of the two estimation equations in our experiments with the text difficulty along with (i) the most correlated feature in our experiment (average word set rankings), (ii) the two text readability measurement formulas of Nguyen and Henkin [11, 17] and their revised version on our experiment corpus, and (iii) the revised version of the formula of Luong, et al. [18]. The correlations of the estimation equations with the text difficulty of the first experiment, the second experiment, and the highest feature (average word set rankings) were 0.95, 0.92 and 0.91, respectively. Hence, while the elimination of strongly correlated features reduced the number of features to be analyzed and minimized processing costs in the text evaluation process, it also lowered the correlation between the estimated equation and the readability of the text. Meanwhile, the experimentation using the two formulas of Nguyen and Henkin [11, 17] on the set of readings in Vietnamese textbooks and Literature in Vietnam at the present yielded the correlation results of only about 0.51 and 0.58, respectively. When we updated the weights of Nguyen and Henkin's formulas [11, 17] and Luong, et al. [18] using our corpus, the correlation with the text readability increased, but it was not as high as the result in our first experiment.

TABLE IV. LIST OF FEATURES HIGHLY CORRELATED WITH THE TEXT READABILITY LEVE

RAW FEATURES			
average word length in syllable	**0.853269**	distinct easy word/distinct words	-0.84908
average word length in character	0.844346	ratio of easy syllables	-0.85065
distinct easy syllables/distinct syllables	0.835926	ratio of easy words	-0.86098
ratio of long sentence (in syllable)	0.818193	ratio of monosyllabic words	-0.86667
ratio of long sentence (in word)	0.809846	ratio of distinct easy syllables	-0.86977
ratio of short sentence (in word)	-0.80448	ratio of distinct easy words	-0.8816
ratio of short sentence (in syllable)	-0.81497		
PART-OF-SPEECH FEATURES			
POS tags/words	-0.8304	adverbs/words	**-0.80988**
ratio of 2-POS tag words	-0.81505	distinct volatile verbs/words	-0.817
ratio of 3-POS tag words	-0.81994	distinct adverbs/words	-0.81554
ratio of multi POS tag words	-0.84525		
SYNTAX-LEVEL FEATURES			
average tree depths (parse tree)	**0.822985**	ratio of simple sentences	-0.81698
BASIC SEMANTIC FEATURES			
semantics/words	-0.82351	ratio of polysemous words	**-0.83606**
ratio of 3-semantic words	-0.82913		
RAW-LEVEL LANGUAGE MODEL FEATURES			
average word set rankings	**0.911331**	average distinct word frequency	-0.83279
average word set frequencies	0.895034	average syllable bigram frequencies	-0.8403
average word list frequencies	0.885074	average frequency of sentence length in word (remove duplicate)	-0.84562
average word rankings	0.863239	average syllable set frequencies	-0.84672
average word trigram frequencies	0.843268	average frequency of sentence length in syllable (remove duplicate)	-0.8502
average syllable trigram frequencies	0.842053	average syllable list frequencies	-0.8535
average syllable set rankings	-0.81599	average character bigram frequencies	-0.86795
average word bigram frequencies	-0.81744	average character trigram frequencies	-0.86852
average syllable bigram rankings	-0.82157	average character bigram rankings	-0.86854
average syllable rankings	-0.82241	average character trigram rankings	-0.86937
average distinct syllable frequency	-0.82974		
POS-LEVEL LANGUAGE MODEL FEATURES			
average word with POS trigram frequencies	**0.846458**	average POS trigram perplexity	-0.82213
average POS bigram perplexity	-0.81658	average POS trigram frequencies	-0.82706
average word with POS bigram frequencies	-0.8171	average POS bigram frequencies	-0.83434
VIETNAMESE-SPECIFIC FEATURES			
distinct borrowed words/distinct words	0.824652	ratio of borrowed words	0.814249
distinct Sino-Vietnamese words/distinct words	0.819849	monosyllabic Sino-Vietnamese words/Sino-Vietnamese words	**-0.83381**

TABLE V. EXPERIMENTAL RESULTS OF THE FIRST REGRESSION ANALYSIS

Intercept	76.76817		
RAW FEATURES			
average word length in character	0.062244	ratio of easy words	-25.0171
average word length in syllable	-4.98138	ratio of long sentence (in syllable)	0.805849
distinct easy syllables/distinct syllables	0.382055	ratio of long sentence (in word)	0.186385
distinct easy word/distinct words	2.023699	ratio of monosyllabic words	-40.62
ratio of distinct easy syllables	-20.0669	ratio of short sentence (in syllable)	8.068247
ratio of distinct easy words	7.342935	ratio of short sentence (in word)	5.744779
ratio of easy syllables	43.08403		
PART-OF-SPEECH FEATURES			
POS tags/words	9.515384	ratio of 2-POS tag words	-4.60083
adverbs/words	-59.9567	ratio of 3-POS tag words	-0.61682
distinct adverbs/words	-25.5114	ratio of multi POS tag words	-16.419
distinct volatile verbs/words	-1.47296		
SYNTAX-LEVEL FEATURES			
average tree depths (parse tree)	0.034768	ratio of simple sentences	-5.72399
BASIC SEMANTIC FEATURES			
ratio of 3-semantic words	19.51943	semantics/words	0.923996
ratio of polysemous words	-21.6204		
RAW-LEVEL LANGUAGE MODEL FEATURES			
average character bigram frequencies	10.45311	average syllable rankings	3.716677
average character bigram rankings	6.02605	average syllable set frequencies	-44.0778
average character trigram frequencies	-3.62059	average syllable set rankings	8.422091
average character trigram rankings	-4.58448	average syllable trigram frequencies	-0.10974
average distinct syllable frequency	0.009618	average word bigram frequencies	38.36505
average distinct word frequency	-0.06197	average word list frequencies	0.070119
average frequency of sentence length in syllable (remove duplicate)	0.005918	average word rankings	4.84E-05
average frequency of sentence length in word (remove duplicate)	-0.00603	average word set frequencies	-3.646
average syllable bigram frequencies	0.208769	average word set rankings	0.002689
average syllable bigram rankings	1.633263	average word trigram frequencies	0.342254
average syllable list frequencies	-22.6624		
POS-LEVEL LANGUAGE MODEL FEATURES			
average POS bigram frequencies	-3.89573	average POS trigram perplexity	-15.3692
average POS bigram perplexity	7.087669	average word with POS bigram frequencies	-14.0184
average POS trigram frequencies	30.56821	average word with POS trigram frequencies	-0.21931
VIETNAMESE-SPECIFIC FEATURES			
distinct borrowed words/distinct words	-0.6103	ratio of borrowed words	0.269794
distinct Sino-Vietnamese words/distinct words	0.232131	monosyllabic Sino-Vietnamese words/Sino-Vietnamese words	-10.2129

TABLE VI. EXPERIMENTAL RESULTS OF THE SECOND REGRESSION ANALYSIS

Intercept	2.808379
volatile verbs/sentences	0.003871
common nouns/words	-73.0814
average word set rankings	0.001179

TABLE VII. CORRELATION COEFFICIENTS OF TWO EXPERIMENTS AND TWO READABILITY FORMULAS OF NGUYEN AND HENKIN [11, 17]

Nguyen and Henkin (1982)	0.51
Nguyen and Henkin (1985)	0.58
Nguyen and Henkin (1982) (revised)	0.85
Nguyen and Henkin (1985) (revised)	0.82
Luong et al. 2018 (revised)	0.87
Only use "average word set rankings"	0.91
Experiment 1	**0.95**
Experiment 2	0.92

2) Machine learning approach

This approach evaluates the role of features in the text classification problem according to the difficulty level. In this study, we used an algorithm called Feature ranking with recursive feature elimination and cross-validated selection of the best number of features (RFECV). Initially, all the features that are examined will be used to classify texts by readability level. The documents will be classified and evaluated by an SVM classification algorithm, using k-fold cross-validation, which splits the corpus into k parts, and then takes k - 1 part for training and the rest part for testing. The features are then removed gradually to test the accuracy of the combination of each feature. Finally, the algorithm evaluates the best combination of documents to classify documents according to their difficulty level. This algorithm has been implemented in the sklean library [32] in Python.

In this experiment, we eliminated from 1 to n-1 number of features, with n being the number of examined features, k = 5, and the evaluation criterion was the classification accuracy. Fig. 1 presents the results of the examination on the number of features and the accuracy achieved through the RFECV algorithm. With about 7 features, the accuracy of the classification process was the highest (85.7%). Table VIII presents the most highly ranked features surveyed by the RFECV algorithm. Out of these 7 features, 6 were raw-level (including 4 language model features), and 1 was Vietnamese-specific feature, with no semantic level features. When compared with the results of 85.17% in the work of Luong, et al. [13] for Vietnamese text, this combination of the seven features achieved slightly higher results with the rate of 85.7%. However, Luong, et al. [13] used some non-standardized text length features, such as numbers of sentences, words, syllables, characters, distinct words, and distinct syllables. These characteristics have proven to be valuable in assessing the difficulty of text in textbooks, when reading time is limited within the framework of a lesson [12]. Therefore, we also conducted an empirical evaluation of the features mentioned in 3.1 together with non-standard text length features. Fig. 2 presents the ranking result and Table IX lists most highly ranked features in this experiment, including a non-standardized feature (number of words), 16 raw-level features, 5 POS-level features, 2 syntax-level features, 9 language model features, 4 Vietnamese-specific features, and no semantic level characteristics. It was possible that the semantic-level features were highly correlated with the readability level but were not suitable for the construction of a readability evaluation model. Another possibility would be that the features examined were too simple or inappropriate with the corpus in question. Other in-depth studies on these characteristics are needed to evaluate these possibilities. Table X presents the accuracy rates of the recent publications of Luong, et al. [12, 13] and of our two experiments on text readability classification on the corpus of Vietnamese and Literature textbooks. With 24 features (including non-standardized length features), the accuracy rate of the classification process was 88.14%, which was higher than those of Luong, et al. [12] and Luong, et al. [13] by 3% to 4%.

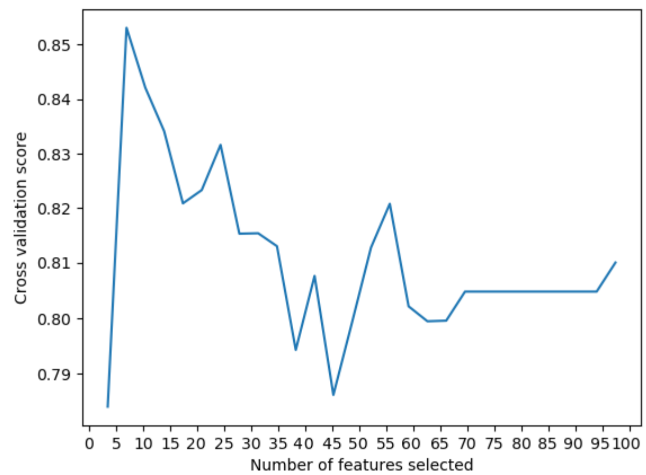

Fig. 1. Experiment Result on the Numbers of Features (without Non-Standardized Length Features).

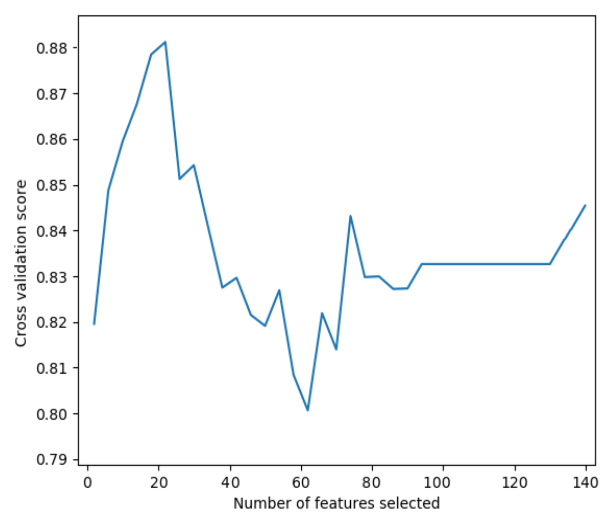

Fig. 2. Experiment Result on the Numbers Of Features (with Non-Standardized Length Features).

TABLE VIII. MOST HIGHLY RANKED FEATURES (WITHOUT NON-STANDARDIZED LENGTH FEATURES)

average word length in syllable
distinct easy syllables/distinct syllables
average word set frequencies
average word list frequencies
average syllable trigram frequencies
average syllable bigram rankings
distinct Sino-Vietnamese words/distinct words

TABLE IX. MOST HIGHLY RANKED FEATURES (WITH NON-STANDARDIZED LENGTH FEATURES)

number of words
average word length in character
ratio of long sentence (in syllable)
ratio of long sentence (in word)
distinct common nouns/distinct words
distinct parallel conjunctions/distinct words
ratio of single POS tag words
adverbs/sentences
average no. distinct conjunction word
average no. conjunction word
average word set frequencies
average word list frequencies
average word trigram frequencies
average syllable trigram frequencies
average word bigram frequencies
average syllable rankings
average syllable set rankings
average syllable bigram rankings
average word with POS trigram frequencies
ratio of borrowed words
ratio of Sino-Vietnamese words
ratio of distinct borrowed words
ratio of distinct Sino-Vietnamese words
polyphonic Sino-Vietnamese words/Sino-Vietnamese words

TABLE X. ACCURACY RATES OF THE TEXT CLASSIFICATION MODELS BY READABILITY, USING 69 SELECTED FEATURES, COMPARED WITH PREVIOUS WORKS

Luong et al. (2017)	84.34
Luong et al. (2018)	85.17
Our experiment (without non-standardized length features)	85.70
Our experiment (with non-standardized length features)	88.14

IV. CONCLUSIONS

In this study, we examined the effects of linguistic features at all levels on the readability assessment of Vietnamese texts. We extracted a total of 271 features from Vietnamese textbooks for primary school students and Literature for middle and high school students in Vietnam to explore. These features range from superficial and word-level features to grammatical and fundamental semantic features. We also surveyed the n-gram features to evaluate the role that the language model plays in determining the difficulty of Vietnamese text.

We conducted the examinations in two main approaches to the readability problem: the statistical approach and the machine learning approach. For the statistical approach, we performed a correlation analysis of 271 features with the difficulty of the surveyed documents and selected 56 highly correlated features, with the correlation coefficient values $\geqslant$ 0.8. Next, we used these 56 features to perform a regression analysis to find the coefficients of the features in the formula to predict the readability of the text. Empirical results indicated that the estimation equation built from these 56 features was highly correlated with the difficulty of the text, with the correlation coefficient of 0.95, significantly higher than previous studies of Nguyen and Henkin [11, 17]. Regarding the machine learning approach, we evaluated the role of features in text classification according to the readability level. The evaluating algorithm used was feature ranking with recursive feature elimination and cross-validated selection of the best number of features (RFECV). This algorithm examined specific combinations in the text classification problem to ranked features, utilizing SVM to model classification and K-fold cross-validation to avoid over-fitting. Experimental results show that, with seven features, most of which were shallow features and language model features, the accuracy of the classification model obtained the highest accuracy (~85.7%). When experimenting with additional non-standardized text length features, the classification results showed a significant improvement over the existing features of Luong et al. [12, 13].

For future works, we will collect additional corpora on different domains to explore the features that would be useful in evaluating the readability of documents in those domains. Deeper features at the semantic level such as coherence and cohesion will also be investigated to detect better combinations for assessing the readability of Vietnamese text.

ACKNOWLEDGMENT

The author(s) received no financial support for the research, authorship, and/or publication of this article.

CONFLICT OF INTEREST

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

REFERENCES

[1] A. Bailin and A. Grafstein, *Readability: Text and Context*. Palgrave Macmillan UK, 2016.

[2] J. S. Chall and E. Dale, *Readability Revisited: The New Dale-Chall Readability Formula*. Northampton, Massachusetts: Brookline Books, 1995.

[3] G. H. Mc Laughlin, "SMOG grading-a new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639-646, 1969.

[4] L. Si and J. Callan, "A Statistical Model for Scientific Readability," in *Proceedings of the Tenth International Conference on Information and Knowledge Management*, New York, NY, USA, 2001, pp. 574-576: ACM.

[5] K. Collins-Thompson and J. Callan, "Predicting Reading Difficulty with Statistical Language Models," *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, no. 13, pp. 1448-1462, November 2005.

[6] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proceedings of the conference on empirical methods in natural language processing*, 2008, pp. 186-195.

[7] 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, 2012, pp. 163-173: Association for Computational Linguistics.

[8] M. Sinha and A. Basu, "A study of readability of texts in Bangla through machine learning approaches," *Education and Information Technologies*, vol. 21, no. 5, pp. 1071-1094, 2016.

[9] S. Vajjala 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*, New Orleans, Louisiana, 2018, pp. 297-304: Association for Computational Linguistics.

[10] M. Al Khalil, H. Saddiki, N. Habash, and L. Alfalasi, "A Leveled Reading Corpus of Modern Standard Arabic," in *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, 2018: European Language Resource Association.

[11] L. T. Nguyen and A. B. Henkin, "A Second Generation Readability Formula for Vietnamese," *Journal of Reading*, vol. 29, no. 3, pp. 219-225, 1985.

[12] 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)*, 2017, pp. 36-41.

[13] A.-V. Luong, 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, pp. 231-236.

[14] 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.

[15] S. Zeno, T. A. S. Associates, R. T. Millard, and R. Duvvuri, *The Educator's Word Frequency Guide*. Touchstone Applied Science Associates, 1995.

[16] J. H. Lee 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.

[17] L. T. Nguyen and A. B. Henkin, "A Readability Formula for Vietnamese," *Journal of Reading*, vol. 26, no. 3, pp. 243-251, 1982.

[18] A.-V. Luong, 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, pp. 198-202.

[19] Đ. T. N. Nguyễn, A.-V. Lương, and Đ. Đinh, "Affection of the part of speech elements in Vietnamese text readability," *Acta Linguistica Asiatica*, vol. 9, no. 1, 01/30 2019.

[20] F. Dell'Orletta, M. Wieling, G. Venturi, A. Cimino, and S. Montemagni, "Assessing the Readability of Sentences: Which Corpora and Features?," in *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, Baltimore, Maryland, 2014, pp. 163-173, references_files/DellOrletta2014.pdf: Association for Computational Linguistics.

[21] O. De Clercq, V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken, "Using the crowd for readability prediction," *Natural Language Engineering*, vol. 20, no. 3, pp. 293-325, 2014.

[22] K. Berendes *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*, vol. 110(4), pp. 518–543, November 2017.

[23] G. Sun, Z. Jiang, Q. Gu, and D. Chen, "Linear model incorporating feature ranking for Chinese documents readability," in *The 9th International Symposium on Chinese Spoken Language Processing*, Singapore, 2014, pp. 29-33, references_files/Sun2014.pdf: IEEE.

[24] Y.-H. Chen and P. Daowadung, "Assessing readability of Thai text using support vector machines," *Maejo International Journal of Science and Technology*, vol. 09, no. 3, pp. 355-369, 2015.

[25] H.-C. Tseng, B. Chen, T.-H. Chang, and Y.-T. Sung, "Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts," *Natural Language Engineering*, vol. 25, no. 3, pp. 331-361, 2019.

[26] *Nghị quyết số 40/2000/NQ-QH10 của Quốc hội Khóa 10 Nước Cộng hòa Xã hội Chủ nghĩa Việt Nam về đổi mới chương trình giáo dục phổ thông (Resolution No. 40/2000/NQ-QH10 of the 10th National Assembly of the Socialist Republic of Vietnam on Renovating the Program for General Education)*, 2000.

[27] M. Vogel and C. Washburne, "An Objective Method of Determining Grade Placement of Children's Reading Material," *The Elementary School Journal*, vol. 28, no. 5, pp. 373-381, 1928.

[28] J. R. Bormuth, "Readability: A New Approach," *Reading Research Quarterly*, vol. 1, no. 3, pp. 79-132, 1966.

[29] P. T. P. Uyen, N. T. H. Tung, T. Thinh, and D. An, "Vietnamese Span-based Constituency Parsing with BERT Embedding," in *The 11th International Conference on Knowledge and Systems Engineering (KSE 2019)*, Da Nang, Vietnam.

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

[31] H. Phe, *Từ điển tiếng Việt (Vietnamese dictionary)*, 8th ed. Da Nang Publishing House, 2017.

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.