EXAMINING THE PART-OF-SPEECH FEATURES IN ASSESSING THE READABILITY OF VIETNAMESE TEXTS

An-Vinh LUONG
Computational Linguistics Center, University of Science, Ho Chi Minh City, Vietnam
anvinhluong@gmail.com

Diep NGUYEN
Department of Linguistics, University of Social Sciences & Humanities, Ho Chi Minh City, Vietnam
nhudiep2004@gmail.com

Dien DINH
Computational Linguistics Center, University of Science, Ho Chi Minh City, Vietnam
ddien@fit.hcmus.edu.vn

Abstract

The readability of the text plays a very important role in selecting appropriate materials for the level of the reader. Text readability in Vietnamese language has received a lot of attention in recent years, however, studies have mainly been limited to simple statistics at the level of a sentence length, word length, *etc.* In this article, we investigate the role of word-level grammatical characteristics in assessing the difficulty of texts in Vietnamese textbooks. We have used machine learning models (for instance, Decision Tree, K-nearest neighbor, Support Vector Machines, *etc.*) to evaluate the accuracy of classifying texts according to readability, using grammatical features in word level along with other statistical characteristics. Empirical results show that the presence of POS-level characteristics increases the accuracy of the classification by 2-4%.

Keywords: text readability; text difficulty; Vietnamese text readability; text classification; school textbooks

Povzetek

Berljivost besedila ima zelo pomembno vlogo pri izbiri ustreznih gradiv za raven bralca. Berljivost besedil v vietnamskem jeziku pridobiva pozornost šele v zadnjih letih in dosedanje študije so omejene na preproste ocene na osnovi statističnih podatkov za dolžino stavka, dolžino besed in podobnih značilnosti. V tem članku raziskujemo vlogo slovničnih značilnosti na besedni ravni pri ocenjevanju težavnosti besedil v vietnamskih učbenikih. Za oceno natančnosti razvrščanja besedila glede na berljivost smo uporabili modele strojnega učenja (na primer drevo odločitve, K-najbližji sosed, podporni vektorski stroji itd.) Empirični rezultati kažejo, da upoštevanje različnih značilnosti na nivoju besednih vrst poveča natančnost klasifikacije za 2-4%.

Ključne besede: berljivost besedila; raven enostavnosti; berljivost vietnamskih tekstov; klasifikacija tekstov; šolski učbeniki

Acta Linguistica Asiatica, 10(2), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.2.127-142

1 Introduction

In today's era of information explosion, thousands of documents with different contents and in different languages get released every second. Such documents have different levels of readability; some are easy to read and understand while others are more difficult and demand larger amount of time and knowledge to get through. It is generally known that the best way to assess whether a text is easy or difficult is to ask readers to read or skim that text, though this can be time-consuming for the readers. Therefore, we suppose that there exists some kind of a method that assists a reader to determine the readability of the text, upon which they make further decision on whether they would continue reading or not. Recently gaining a lot of focus is readability, which is one of such methods.

Brown and colleagues state that

> readability is a concept that describes the degree to which a text is easy or difficult to read. A readability index is a numerical scale that estimates the readability or degree reading difficulty that native speakers are likely to have in reading a particular text. (Brown et al., 2012).

Determining the readability index of a document is to determine how difficult the text is, which gives a reader information on whether the document is suitable for them to read to understand it in a reasonable amount of time. Information on readability is useful in many different fields of science as well as in everyday life. It can be used when assisting scientists at publishing articles, helping text editors (writers, journalists, *etc.*) to create documents suitable specific audience, or else for manufacturers to produce readable manuals. Above all, information on readability is of most importance in education, especially in second language education. It is used when textbooks are compiled, or for educators to make decisions on appropriate texts are made.

Research on the difficulty of texts originates back to the late 19th century when Lucius Adelno Sherman wrote that "the average length of sentences has been decreasing over time" (Sherman, 1893). Many books on readability have been published since, however, they mostly applied for English, such the work of Dale and Chall (1948), Si and Callan (2001), Schwarm and Ostendorf (2005), Chall and Dale (1995), Chen and Meurers (2018), etc., and some languages that were treated as lingua franca at some point in the history or some part of the world. Thus we find works on French (François (2014), François & Fairon (2012), *etc.*), Chinese (Chen et al. (2013); Jiang et al. (2018); Sun et al. (2014), *etc.*), Spanish (Coco et al. (2017); I. Parkeret al. (2001); Spaulding (1956), etc.), Arabic (Al-Ajlan et al. (2008); Al-Tamimi et al. (2014); Al Khalil et al. (2018); Saddiki et al. (2015); Saddiki et al. (2018), *etc.*), and other languages.

For less-resource languages, studies on the readability of texts are still limited, and Vietnamese is one of such languages. In Vietnamese, there some publications that date

back to 1980ies (Nguyen & Henkin 1982, 1985), and recent studies of Luong et al. (2017, 2018a, 2018b), Điệp (2019) and Luong & Tran (2019). These studies have shown some valuable features for assessing text readability in Vietnamese, but the results are limited and further research on the topic is necessary.

In this research, we examine features on the level of parts of speech (POS-level) and assess the readability of literary texts based on them. The texts were taken from literary textbooks for school students from grade 2 to grade 12, which corresponds to the students' age 7 to age 17. This method inherits the results of the Luong et al. (2017, 2018b) with the addition of a number of grammatical features at word-level to build a text-based classifier based on readability through some machine learning methods like Decision Tree, K-nearest neighbor, Support Vector Machines, *etc.*

The article is thus organized as follows. Section 2 presents some ground works on text readability and previous literature on Vietnamese text readability. It also introduces the features that we surveyed and used in this study to develop models for assessing the readability of Vietnamese texts, using some classification algorithms along with experimental results. Section 0 presents the results of the study and discusses them, and the final Section 5 offers an overall conclusion to the topic.

2 Related works

2.1 Different approaches in previous studies

Previous studies of text readability can be grouped into two groups based on either they undertake traditional approach or corpus-based approach.

Traditional approach uses conventional statistical methods on the documents to select high correlation factors with the readability of texts and then use regression analysis to create formulas for measuring the readability. The factors examined are typically shallow features, also called easy-to-extract features, such as average sentence length, average word length, percentage of difficult words in the documents. Representative researches with this approach produces the Dale-Chall formula (Dale & Chall, 1948), the Gunning Fog Index (Robert, 1952), the SMOG formula (Mc Laughlin, 1969), the Flesch-Kincaid grade level readability (Kincaid, Fishburne, Rogers, & Chissom, 1975), the new Dale-Chall formula (Chall & Dale, 1995), and others.

On the other hand, corpus-based approach approach has been developed in recent years due to the fast development computer science and machine learning algorithms. Studies in this approach see the problem of assessing the readability of text as a classification problem, and use machine learning models to classify text by layer of readability based on extracted features. Representative works are those of Si & Callan (2001); Collins-Thompson & Callan (2005); Schwarm & Ostendorf (2005); Heilman et al.

(2007); Pitler & Nenkova (2008); Feng et al. (2010); Vajjala & Meurers (2012); Jiang et al. (2015); Wang & Andersen (2016); Chen & Meurers (2018), and others.

2.2 Studies on text readability in Vietnamese

The research on the readability of the text in Vietnamese is still quite small and their results are limited. Nguyen et al., have introduced two formulas to measure the readability of Vietnamese texts (Nguyen & Henkin, 1982, 1985). These two formulas base on features such as the average length of sentences or words, and the ratio of difficult words in texts. The weak point in these works is that the two formulas were surveyed and evaluated on a relatively small amount of data; on 20 documents in Nguyen & Henkin, 1982 and 54 documents in Nguyen & Henkin, 1985.

Luong et al. (2017) conducted a survey of texts extracted from literary textbooks for Vietnamese high school students and suggested to use the feature of text length to classify texts according to readability. Experimental results show that the length of texts has a great influence on the classification results, and is to be used to evaluate texts in Vietnamese textbooks.

Luong et al. (2018a) introduced a new formula for measuring the readability of Vietnamese texts. This formula is based on a survey of 1,200 documents classified into 3 levels of difficulty (easy, medium and difficult). The features of the average length of the sentence, of the word and the ratio of difficult words in the text have been chosen formulas criteria.

In addition, Luong et al. (2018b) published another study on the readability of texts using the proportional features of proper nouns and Vietnamese specific characteristics such as Sino-Vietnamese words ratio, borrowed words ratio, and dialect words ratio within documents. Experimental results show the contribution of these features in improving the accuracy of classification processes.

Điệp et al. (2019) presented the statistical analyses on the frequency of POS tags in Vietnamese texts. They conducted a survey of 209 texts extracted from Vietnamese textbooks from grade 2 to grade 5 (corresponding to age 7 to 10) for primary students according to the general curriculum in Vietnam. Their results showed that words such as common nouns and volatile verbs were common in the examined documents. In addition, through the correlation analysis between the frequencies of POS tags and the readability of the surveyed documents, they also proved a high correlation between the ratio of common nouns and the ratio of prepositions with the readability level in the examined texts.

Furthermore, Luong & Tran (2019) introduced a method of evaluating the readability of documents by comparing difficulty correlation between different documents. They built a set of 30 texts – which were graded the readability level – as the standard.

The documents in this research will be compared to the standard texts proposed by Luong & Tran (2019) to determine the readability level.

2.3 Features

In this section, we will introduce features that used to classify texts with readability level. These features include the so-called traditional or grammatical features at the word level (features (1) – (4)), and features (5) – (10) proposed by Luong et al. (2017, 2018b) that have been defined relevant in Vietnamese literary texts and are the focus of this research.

(1) Average sentence length. The average length of sentences is a common factor in most studies of text readability. The length of sentences is very important in the process of documenting and reading texts. If a text has too many long sentences, it may make it difficult for the reader to fully understand meanings of its long sentences. On the other hand, using only short sentences may make the text discrete and incoherent, which could make the reader experience difficulties. Therefore, the length of sentences is a very important factor in assessing the readability of text. The average sentence length features commonly used are average sentence length in words (ASLW), in syllables (ASLS), and in characters (ASLC).

(2) Average word length. Carver (1976) showed a linear correlation between the length of words and the readability of text, which is a commonly used factor in studies on text readability. The average word length in syllables (AWLS) and in characters (AWLC) are commonly used with length features.

(3) Percentage of difficult words. In many studies, the percentage of difficult words is one of the most valuable factors when assessing the readability of texts. These studies often use a list of easy vs. difficult words in a language as the base for calculation. However, building such a list takes a lot of effort, and therefore many studies used statistical lists of words according to their frequency of use instead; they chose most commonly used words in a particular language and treated them as a list of easy words.

In this study, we use 3,000 most popular words extracted from the statistical list of words in the Vietnamese texts of the researching group of Dinh et al. (Dinh, Nguyen, & Ho, 2018) as the basis for calculating difficult word rate. The features we surveyed are Percentage of Difficult Words (PDW), Percentage of Unique Difficult Words (PDDW).

(4) Percentage of difficult syllables. Vietnamese writing is monosyllabic in nature. Every “syllable” is written as though it were a separate dictation-unit with a space before and after. Such a unit is called morphosyllable or “tiếng” in Vietnamese. Each morphosyllable tends to have its own meaning and consequently a strong identity. However, these morphosyllables are not automatically combined into ‘words’ as the linguistic notion of word commonly applies for European languages (Tran et el., 2007),

which leads to difficulties for readers, especially those with low reading skills, to distinguish the boundaries between words.

For this reason, we consider syllables as an important language unit of Vietnamese to make statistics and use as a characteristic for examination. In this work, we use 3,000 most popular syllables of Dinh et al. (2018) to extract the following two features: percentage of difficult syllables (PDS) and percentage of distinct difficult syllables (PDDS).

(5) Text length features. In the study by Luong et al. (2017), results showed the essential role of the text length features in assessing readability. The features that Luong et al. surveyed and are relevant to this study are as follows. The total number of sentences (NSen), the total number of words (NWo), the total number of syllables (NSyl), the total number of characters (NCha), the total number of distinct words (NDWo), and finally the total number of distinct syllables (NDSyl).

In the article published in 2018, Luong et al. introduced some additional features for assessing readability of Vietnamese texts:

(6) Percentage of Sino-Vietnamese words. Vietnam has spent more than 1,000 years of domination by the Chinese feudal dynasties (111BC - 905AC). During that period, the Vietnamese language was strongly influenced by Chinese culture and language, and those influences continue to this day. Vocabulary of the Vietnamese language consists of more than 60% of words of Chinese-origin, the called as Sino-Vietnamese words (DeFrancis, 1977). These Sino-Vietnamese words are often used in the official and ceremonial language and are therefore considered as more difficult compared to originally Vietnamese words of the same meaning.

In this study, we examined features such as percentage of Sino-Vietnamese Words (PSVW), percentage of distinct Sino-Vietnamese words (PDSVW), and the proportion of distinct Sino-Vietnamese within all distinct words (DSVW/DW).

(7) Percentage of borrowed words. Similar to Sino-Vietnamese words, many words from other languages entered Vietnamese. This foreign influence was especially strong during the French invasion of Vietnam in the middle of the 19th century. Such words of French, English and other origin undertook Vietnamese phonetic transcriptions (Alves, 2009) and are nowadays used in the official and scientific language. It is estimated that they influence the readability and are therefore taken into account as the percentage of borrowed words (PBW), the percentage of distinct borrowed words (PDBW), and the proportion of distinct borrowed words within all distinct words (DBW/DW).

(8) Percentage of dialectal words. Vietnamese territory has many different regions with different cultural and linguistic characteristics. Regions tend to localize general Vietnamese language and use it as their own regional language. Consequently, such dialectal vocabulary, which is not available in the standard Vietnamese language, is

thought to be difficult for readers. In assessing text readability the feature appears either as the percentage of dialect words (PDiaW), the percentage of distinct dialect words (PDDiaW), or the proportion of distinct dialect words within all distinct words (DDiaW/DW).

(9) Percentage of proper nouns. According to Luong et al. (2018b), the more proper nouns in the text, the more effort the reader will have to memorize those objects, and therefore the text is considered more difficult. For this reason we have decided to take into account the characteristics of proper nouns in this experiment. The features are defines in the following way. Nr/Sen is the abbreviation for the proportion of proper nouns within sentences. Nr/W is the abbreviation for the number of proper nouns in comparison to all words. Nr/DW stands for the number of proper nouns that is divided by the number of distinct words. DNr/Sen points to the proportion of distinct proper nouns within the overall number of sentences. DNr/W stands for the number of distinct proper nouns divided by the number of words. Finally, DNr/DW is the abbreviation for the number of distinct proper nouns divided by the number of distinct words.

(10) Other parts of speech and their elements. In this study, we also used other POS tags such as countable noun, directional verb, parallel association, *etc.* to experiment with the model. Table 1 is a list of tags used. These POS tags are derived from the CLC_VN_Toolkit tool, which has been developed by the Computational Linguistics Center, Ho Chi Minh City University of Science[1]. This is a tool for pre-processing, sentences segmentation, words segmentation, part-of-speech tagging (POS), named entity labeling, *etc.* for Vietnamese texts. Similar to proper nouns, we use features with abbreviated symbols for each POS element. The number POS divided by the number of sentences is POS/Sen. The number of POS divided by the number of words is POS/Wo. The number of POS divided by the number of distinct words is POS/DWo. The number of distinct POS divided by the number of sentences is DPOS/Sen. The number of Distinct POS divided by the number of words is DPOS/Wo. The number of Distinct POS divided by the number of distinct words is DPOS/DWo. The abbreviation 'POS' is generally replaced by POS tags as shown in Table 1 except for proper nouns (Nr) already presented in the work of Luong et al. (2018b).

Table 1: List of Vietnamese POS tags used in CLC_VN_Toolkit

POS	Tag	POS	Tag
Countable nouns	Nc	Quality adjectives	Aa
Concrete nouns	Nu	Demonstrative pronouns	Pd
Temporal nouns	Nt	Personal pronouns	Pp

[1] CLC website: http://clc.hcmus.edu.vn

POS	Tag	POS	Tag
Numerals	Nq	Adverbs	R
Common nouns	Nn	Prepositions	Cm
Proper nouns	Nr	Parallel conjunctions	Cp
Directional verbs	Vd	Subordinating conjunctions	Cs
State verbs	Ve	Modifiers	M
Comparative verbs	Vc	Emotion words	E
Volatile verbs	Vv	Foreign words	FW
Directional co-verb	D	Onomatopoeia	ON
Quantity Adjectives	An	Idioms	ID

3 Experiment

In this study, we used the corpus of 371 literary texts of Luong et al. (2018b) for experimentation. These documents were taken from Vietnamese textbooks for primary, middle and high school students in Vietnam. We divided the texts into groups based on:

(1) grade level (from grade 2 to grade 12);
(2) level of education (Primary, Middle and High school).

Table 2 presents the basic statistics of the corpus. The features mentioned in Section 2.3 are used to build the classification models for text readability.

Table 2: The statistics of the corpus of 371 literary documents of Luong et al. (2018b)

Grade	2	3	4	5	6	7	8	9	10	11	12
Number of texts	67	62	40	40	28	13	17	21	15	19	49
Average number of sentences	18.34	19.63	21.53	21.43	54.75	46.38	65.76	107.33	60.67	105.16	111.65
Average number of words	158.06	192.31	231.28	244.4	679.54	676.92	969.24	1447.4	861.73	1359.9	1710.3
Average number of distinct words	100.63	125.58	144.3	152.78	304.86	329.69	394.29	526.29	368.4	510	576
Average number of syllables	178.48	221.98	276.1	288	784.11	820.85	1131.5	1709.7	1006.5	1579.1	2179.4
Average number of distinct syllables	111.36	141.53	164.78	173.35	327.54	372.46	428.35	555.52	390.07	534.95	594.2
Average number of characters	826.8	1065.4	1335	1395.9	3709	3942.3	5401.9	8160	4860	7535.1	10761

Grade	2	3	4	5	6	7	8	9	10	11	12
Average sentence length in words	9.14	10.61	11.59	12.69	14.01	17.99	17.78	18.23	15.04	15.67	16.68
Average sentence length in syllables	10.36	12.34	14.08	15.21	16.14	22.3	21.34	22.07	17.72	18.72	22.17
Average sentence length in characters	48.3	59.57	68.61	74.37	76.67	108.69	103.38	106.62	85.79	90.66	111.21
Average word length in syllables	1.13	1.16	1.2	1.19	1.15	1.23	1.19	1.2	1.17	1.18	1.32
Average word length in characters	5.25	5.61	5.84	5.77	5.43	5.98	5.74	5.78	5.67	5.7	6.59

We conducted experiments by using several classification algorithms such as Decision Tree (denoted as D-TREE), K-nearest neighbor (K-NN), Multi-layer Perceptron (MLP), Random Forest (RND-FRST), and Support Vector Machines (SVM). In this study, we used Scikit-learn, a machine-learning library for the python programming language for the experiments. With D-TREE and RND-FRST, we used two common impurity measures: Entropy and Gini index. In order to avoid overfitting we used k-fold cross validation during training and testing; randomly dividing the corpus into 5 parts (4 parts for training and 1 for testing). The best features combinations of Luong et al. (2017) and Luong et al. (2018b) are used as the baselines for the experimental process. Tables 3 and 4 show the best practices on 4 metrics: accuracy (Acc), precision (P), recall (R), and F1-score (F1).

Table 3: Classification results performed on grade-level documents

Feature set	Acc	P	R	F1
	D-TREE (ENTROPY)			
Luong2017	0.3828	0.2893	0.2772	0.2728
Luong2018	0.4206	0.3488	0.3276	0.3142
Luong2017 + PSVW, Nq/Sen, DMSen	**0.4449**	**0.3749**	**0.364**	**0.3552**
Luong2017 + PSVW, Aa/Sen, DCm/Wo	0.4341	0.3714	0.3523	0.3492
Luong2017 + PSVW, DNr/DWo, Cp/DWo	0.4204	0.3709	0.3467	0.3359
	D-TREE (GINI)			
Luong2017	0.3855	0.3049	0.2973	0.289
Luong2018	0.3909	0.3038	0.2959	0.2888
Luong2017 + PSVW, Aa/Sen, Cm/Wo	0.4448	0.3506	**0.3829**	**0.3538**
Luong2017 + PSVW, DNr/DWo, Nq/Wo, DAa/Sen, Nq/DWo, Cm/Wo	0.4368	**0.3849**	0.3472	0.3409
Luong2017 + PSVW, Nq/Wo, DCm/Sen	**0.4502**	0.3649	0.3473	0.3375

Feature set	Acc	P	R	F1
		KNN		
Luong2017	0.4556	0.2996	0.3097	0.2928
Luong2018	0.4475	0.2929	0.3038	0.2877
Luong2017 + PSVW, Aa/Sen, Cm/Sen	**0.4609**	**0.3236**	**0.3283**	**0.3069**
Luong2017 + PSVW, Nr/Sen, Cm/Sen	0.4584	0.3075	0.3232	0.3035
Luong2017 + PSVW, Cp/Sen, Cm/DWo	0.4476	0.3134	0.3182	0.3025
		MLP		
Luong2017	0.3882	0.2916	0.3034	0.2696
Luong2018	0.3883	0.3246	0.2845	0.2724
Luong2017 + PSVW, DNr/DWo, DAa/Sen, Cm/Wo, DNr/Sen	0.4286	0.3258	0.3404	**0.3058**
Luong2017 + PSVW, DNr/DWo, DCp/Sen	**0.4421**	0.3128	**0.3447**	0.2993
Luong2017 + PSVW, Aa/Sen, DD/Sen	0.38	**0.3488**	0.3098	0.2955
		RND-FRST (ENTROPY)		
Luong2017	0.4529	0.3569	0.3577	0.3403
Luong2018	0.4689	0.3952	0.3503	0.3477
Luong2017 + PSVW, DNr/DWo, Aa/DWo	**0.5041**	0.4291	**0.4029**	**0.3897**
Luong2017 + PSVW, Nr/Sen, DM/Wo	0.4772	**0.4629**	0.382	0.3889
Luong2017 + PSVW, Nn/Sen, Aa/DWo	0.4826	0.4178	0.4011	0.3811
		RND-FRST (GINI)		
Luong2017	0.4392	0.3191	0.3206	0.3071
Luong2018	0.4636	0.345	0.3365	0.3195
Luong2017 + PSVW, Nr/Sen, Cp/Wo	**0.523**	**0.4256**	**0.4089**	**0.4051**
Luong2017 + PSVW, DNr/DWo, DPp/Sen	0.4989	0.402	0.3883	0.3766
Luong2017 + PSVW, Nq/Sen, Nn/Sen	0.4852	0.4078	0.3809	0.371
		SVM (LINEAR)		
Luong2017	0.4446	0.3402	0.3257	0.3177
Luong2018	0.477	0.3892	0.3611	0.3538
Luong2017 + PSVW, DNr/DWo, Nq/Wo, DAa/Sen, Cm/Wo	**0.5148**	**0.4657**	**0.4219**	**0.418**
Luong2017 + PSVW, DNr/DWo, Nq/Wo, Nq/DWo, Cm/Wo	0.5068	0.4479	0.4099	0.4134
Luong2017 + PSVW, DNr/DWo, Nq/Wo	0.5069	0.4464	0.4132	0.408

Table 4: Classification results performed on school-level documents

Feature set	Acc	P	R	F1
		D-TREE ENTROPY		
Luong2017	0.7845	0.7268	0.7006	0.7021
Luong2018	0.8167	0.7594	0.7489	0.7511
Luong2018 + Nc/DWo, DPp/DWo	**0.8329**	**0.7881**	**0.7729**	**0.7761**
Luong2018 + DVc/Sen, DID/Wo	0.8221	0.7792	0.7584	0.7623
Luong2018 + FW/Wo, DNn/Sen	0.8221	0.7739	0.7569	0.7607
		D-TREE GINI		
Luong2017	0.7925	0.7234	0.6985	0.7008
Luong2018	0.7925	0.7174	0.7033	0.7049
Luong2018 + Nu/Sen, D/Wo	**0.8169**	0.7531	**0.7429**	**0.743**
Luong2018 + D/Sen, DNn/Sen	0.8087	0.7568	0.7341	0.7322
Luong2018 + Nc/Sen, DCp/DWo	0.8114	0.7441	0.7338	0.7316
		KNN		
Luong2017	0.7708	0.6687	0.656	0.6594
Luong2018	0.7708	0.6687	0.656	0.6594
Luong2018 + Vv/Wo, DVv/Sen	**0.7815**	**0.6846**	**0.6688**	**0.6746**
Luong2018 + Aa/Sen, DNu/Wo	0.7762	0.6759	0.6612	0.6655
Luong2018 + Aa/Sen	0.7762	0.6759	0.6612	0.6655
		MLP		
Luong2017	0.6589	0.4973	0.5855	0.5169
Luong2018	0.6846	0.5701	0.631	0.5666
Luong2018 + Nr/Wo, DPp/Wo	**0.7954**	0.7124	**0.7029**	**0.6723**
Luong2018 + Aa/Sen, FW/Wo	0.7707	**0.7555**	0.7005	0.6652
		RND-FRST ENTROPY		
Luong2017	0.8221	0.757	0.7368	0.7367
Luong2018	0.8355	0.7743	0.7547	0.7596
Luong2018 + M/Wo, DNq/Sen	**0.8599**	0.8138	**0.7956**	**0.802**
Luong2018 + Nc/Sen, Nq/Sen	0.8544	0.8126	0.789	0.7939
Luong2018 + Nq/Sen, Aa/Sen	0.8571	**0.8169**	0.7824	0.7903
		RND-FRST GINI		
Luong2017	0.8222	0.7569	0.7411	0.7439
Luong2018	0.8302	0.7735	0.7528	0.7573
Luong2018 + Nq/Wo, DON/DWo	**0.8653**	**0.8182**	**0.8004**	**0.8062**
Luong2018 + Nc/Sen, Nq/Sen	0.8652	0.8173	0.7973	0.8031
Luong2018 + Nq/Wo	0.8491	0.7978	0.7837	0.7879

Feature set	Acc	P	R	F1
	SVM LINEAR			
Luong2017	0.8274	0.785	0.7626	0.7644
Luong2018	0.8517	0.8107	0.7842	0.7903
Luong2018 + Aa/Sen, DPp/Wo	**0.8787**	**0.8462**	**0.8206**	**0.8231**
Luong2018 + Aa/Sen, DFW/Wo	0.8733	0.8326	0.8153	0.8182
Luong2018 + D/Sen, DNn/Sen	0.8706	0.833	0.8163	0.8162

From the results presented in Table 3 and Table 4 we can see that, when adding POS features, some features have helped improve the performance of the classification model.

With the experiments in grade-level grouping, accuracy increased from the value 0.4770 of the work of Luong2017 to the value 0.5148 when adding the features PSVW, DNr/DWo, Nq/Wo, DAa/Sen, Cm/Wo with the SVM classifier. Similarly, precision, recall and F1-score also increased from 0.3892, 0.361, and 0.3538 respectively in Luong2017 to 0.4657, 0.4219, and 0.4180 respectively with the SVM classifier. In experimental results, the most accurate features combination is the combination (Luong2017 + PSVW, Nr/Sen, Cp/Wo), implemented on the Random Forest classifier (Gini index). However, the combination that yield the highest precision and F1-score is the combination (Luong2017 + PSVW, DNr/DWo, Nq/Wo, DAa/Sen, Cm/Wo). Among the POS features surveyed, the feature DNr/Dwo (Number of Distinct Proper Nouns divided by number of Distinct Words) feature appears the most in high performing experiments (appears 9 times in Table 3). This shows that the DNr/Dwo feature is a good feature for evaluating the readability of Vietnamese texts Besides, some other POS features also appear several times in the Table 3, such as Cm/Wo (5 times), Nq/Wo (5 times), Aa/Sen (4 times), etc. These POS features are also valuable for classifying Vietnamese texts according to difficulty level.

With school-level grouping, the highest experimental results belong to the feature combination (Luong2018 + Aa/Sen, DPp/Wo), implemented on the SVM classifier: the Accuracy, Precision, Recall and F1-score increased from 0.8517 (Luong2018) to 0.8787; from 0.8107 (Luong2018) to 0.8462; from 0.7842 (Luong2018) to 0.8206; and from 0.7903 (Luong2018) to 0.8231 respectively. The feature Aa/Sen (Number of Quality Adjectives divided by number of Sentences) appears the most (6 times) in the Table 4, therefore, this is a valuable feature for assessing the readability of Vietnamese texts. Similarly, features like DNn/Sen (appears 3 times), Nc/Sen (appears 3 times) or Nq/Sen (appears 3 times) are also good features for automatic classification of Vietnamese texts according to the difficulty level.

Experimental results also show that SVM classifier performs best on overall Accuracy, Precision, Recall and F1-score for most feature sets on both school and grade-level. The Random Forest classifier (Gini impurity) archives the best accuracy in

grade-level with the feature set of (Luong2017 + PSVW, Nr/Sen, Cp/Wo). The other classifiers do not seem suitable for the problem of evaluating the readability of Vietnamese text.

4 Discussion and conclusion

Text readability is an important factor affecting the selection and understanding of documents. Numerous studies on text readability have been conducted for English and some other resource-rich languages, while for Vietnamese research results are rare and limited. In this study, we investigated the role of word-level grammatical characteristics in assessing the difficulty of texts in Vietnamese textbooks. We conducted empirical assessments of text readability in 371 literary texts extracted from Vietnamese textbooks primary school students and the literary textbooks for middle and high school students in Vietnam. Some machine learning algorithms for automatic text classification like Decision Tree, K-nearest neighbor, Support Vector Machines, *etc.* were used to classify the texts.

The experimental results presented in Table 3 and Table 4 show that some POS features such as DNr/Dwo, Cm/Wo, Nq/Wo, or Aa/Sen also contribute to the efficiency of classification. Comparing the results to the Luong 2017 results we can conclude that, he feature set (DNr/DWo, Nq/Wo, DAa/Sen, Cm/Wo), and the feature PSVW help increase precision value with SVM classifier in case of the group-by-grade-level corpus. On the other hand, the case of the group-by-school-level corpus, the feature set (Aa/Sen, DPp/Wo) helped the classification process to achieve the highest results for all measurements.

Experiments in this study only used those machine learning classification algorithms that assess whether a feature is valuable for the classification or not. For that reason it is not possible to discuss the potential influence that increasing or decreasing the use of a certain POS would have on the difficulty of the text. Such studies on the correlation of the extracted features with the text readability level are planned to be conducted in the upcoming investigations.

For the future works, we will proceed to collect additional corpora on different domains to look for features that could be useful for evaluating the readability of texts in the responding domains. Deeper features such as sentence-level grammar (syntax, coherence, cohesion, and others) should also be surveyed to find a better combination of features for assessing the readability of Vietnamese texts.

References

Al-Tamimi, A. K., Jaradat, M., Aljarrah, N., & Ghanim, S. (2014). AARI: Automatic Arabic readability index. *International Arab Journal of Information Technology, 11*(4), 370-378.

Alves, M. J. (2009). Loanwords in Vietnamese. *Loanwords in the world's language: A Comparative Handbook*, 617-637.

Brown, J. D., Janssen, G., Trace, J., & Kozhevnikova, L. (2012). A preliminary study of cloze procedure as a tool for estimating English readability for Russian students. In *Second Language Studies Paper* (pp. 1-22): University of Hawai'i at Manoa.

Carver, R. P. (1976). Word Length, Prose Difficulty, and Reading Rate. *Journal of Reading Behavior, 8*(2), 193-203.

Chall, J. S., & Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Northampton, Massachusetts: Brookline Books.

Chen, X., & Meurers, D. (2018). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading, 41*(3), 486-510.

Chen, Y.-T., Chen, Y.-H., & Cheng, Y.-C. (2013). Assessing Chinese Readability using Term Frequency and Lexical Chain. *IJCLCLP, 18*(2), 1-18.

Coco, L., Colina, S., Atcherson, S. R., & Marrone, N. (2017). Readability Level of Spanish-Language Patient-Reported Outcome Measures in Audiology and Otolaryngology. *American journal of audiology, 26*(3), 309-317. doi:10.1044/2017_AJA-17-0018

Collins-Thompson, K., & Callan, J. (2005). Predicting Reading Difficulty with Statistical Language Models. *J. Am. Soc. Inf. Sci. Technol., 56*(13), 1448-1462.

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, 11-28.

DeFrancis, J. (1977). Colonialism and language policy in Viet Nam. The Hague: Mouton.

Dinh, D., Nguyen, T. N., & Ho, H. T. (2018). Building a corpus-based frequency dictionary of Vietnamese. In (pp. 72-98).

Nguyễn Điệp T. N., Lươnga.-V., & Đinh Điền. (2019). Affection of the part of speech elements in Vietnamese text readability. *Acta Linguistica Asiatica*, *9*(1), 105-118. https://doi.org/10.4312/ala.9.1.105-118

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. e., mie. (2010). *A Comparison of Features for Automatic Readability Assessment.* Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Stroudsburg, PA, USA.

François, T., & Fairon, C. (2012). *An AI readability formula for French as a foreign language.* Paper presented at the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill Book Co.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007, April). *Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts.* Paper presented at the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Rochester, New York.

Parker, R. I., Hasbrouck, J. E., & Weaver, L. R. (2001). Spanish readability formulas for elementary-level texts: A validation study. *Reading & Writing Quarterly*, 17(4), 307-322. doi:10.1080/105735601317095052

Jiang, Z., Sun, G., Gu, Q., Yu, L., & Chen, D. (2015). *An Extended Graph-Based Label Propagation Method for Readability Assessment.* Paper presented at the Web Technologies and Applications, Cham.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical Training, Research B*(February), 49.

Luong, A.-V., Nguyen, D., & Dinh, D. (2018a, November). *A New Formula for Vietnamese Text Readability Assessment.* Paper presented at the 2018 10th International Conference on Knowledge and Systems Engineering (KSE).

Luong, A.-V., Nguyen, D., & Dinh, D. (2018b, November). *Assessing the Readability of Literary Texts in Vietnamese Textbooks.* Paper presented at the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS).

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading, 12*(8), 639-646.

Nguyen, L. T., & Henkin, A. B. (1982). A Readability Formula for Vietnamese. *Journal of Reading, 26*(3), 243-251.

Nguyen, L. T., & Henkin, A. B. (1985). A Second Generation Readability Formula for Vietnamese. *Journal of Reading, 29*(3), 219-225.

Pitler, E., & Nenkova, A. (2008). *Revisiting readability: A unified framework for predicting text quality.* Paper presented at the Proceedings of the conference on empirical methods in natural language processing.

Saddiki, H., Bouzoubaa, K., & Cavalli-Sforza, V. (2015). *Text readability for Arabic as a foreign language.* Paper presented at the Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of.

Saddiki, H., Habash, N., Cavalli-Sforza, V., & Al Khalil, M. (2018, July). *Feature Optimization for Predicting Readability of Arabic L1 and L2.* Paper presented at the Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, Melbourne, Australia.

Schwarm, S. E., & Ostendorf, M. (2005). *Reading Level Assessment Using Support Vector Machines and Statistical Language Models.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA.

Sherman, L. A. (1893). *Analytics of literature: a manual for the objective study of English prose and poetry*. Boston, England: Ginn.

Si, L., & Callan, J. (2001). *A Statistical Model for Scientific Readability.* Paper presented at the Proceedings of the Tenth International Conference on Information and Knowledge Management, New York, NY, USA.

Spaulding, S. (1956). A Spanish Readability Formula. *The Modern Language Journal, 40*(8), 433-441. doi:10.1111/j.1540-4781.1956.tb02145.x

Tran, T., Pham, T., Ngo, H., Dien, D., & Collier, N. (2007). Named Entity Recognition in Vietnamese documents. *Progress in Informatics*, 5-13. doi:10.2201/NiiPi.2007.4.2

Al-Ajlan, A. A., Al-Khalifa, H. S., & Al-Salman, A. S. (2008, November). *Towards the development of an automatic readability measurements for arabic language.* Paper presented at the 2008 Third International Conference on Digital Information Management, University of East London, London, UK.

Vajjala, S., & Meurers, D. (2012, June). *On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition.* Paper presented at the Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Montréal, Canada.

Sun, G., Jiang, Z., Gu, Q., & Chen, D. (2014, September). *Linear model incorporating feature ranking for Chinese documents readability.* Paper presented at the The 9th International Symposium on Chinese Spoken Language Processing, Singapore.

François, T. (2014, November). *An analysis of a French as a Foreign Language Corpus for Readability Assessment.* Paper presented at the Proceedings of the third workshop on NLP for computer-assisted language learning, Uppsala, Sweden.

Wang, S., & Andersen, E. (2016, December). *Grammatical Templates: Improving Text Difficulty Evaluation for Language Learners.* Paper presented at the Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan.

Luong, A.-V., Nguyen, D., & Dinh, D. (2017, October). *Examining the text-length factor in evaluating the readability of literary texts in Vietnamese textbooks.* Paper presented at the 2017 9th International Conference on Knowledge and Systems Engineering (KSE).

Al Khalil, M., Saddiki, H., Habash, N., & Alfalasi, L. (2018, May). *A Leveled Reading Corpus of Modern Standard Arabic.* Paper presented at the Proceedings of the 11th Language Resources and Evaluation Conference, Miyazaki, Japan.

Jiang, Z., Gu, Q., Yin, Y., & Chen, D. (2018, August). *Enriching Word Embeddings with Domain Knowledge for Readability Assessment.* Paper presented at the Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA.

Luong, A.-V., & Tran, P. (2019, November). *Assessing the Readability of Vietnamese Texts Through Comparison.* Paper presented at the Computational Data and Social Networks, Cham.