SCHOLARS: Jurnal Sosial Humaniora dan Pendidikan
ISSN: 3047-8618 (Online) | DOI: 10.31959/js.v3i1.2965
Volume 3 , No. 1, June 2025 (pp. 1-16)
Homepage: https://ejournal-polnam.ac.id/index.php/JS/index

Assessment of Readability of Vietnamese Text

Hien Pham[1], Thi-Minh-Huyen Nguyen[2], Phuong Le-Hong[3]*

[1]School of Languages and Tourism, Hanoi University of Industry, Vietnam
[2,3]Vietnam National University, Hanoi
*Corresponding author email: phuonglh@vnu.edu.vn

Abstract: This paper presents a study that has systematically considered the impact of linguistic features on readability for Vietnamese texts. We used a set of selected texts from the primary school, junior high school and high school Vietnamese textbooks as our input data set. The set of linguistic features accounted for various aspects, such as morphological features, part-of-speech features, syntactic features and discourse features. Our study targets qualitatively and quantitatively developing a readability scale for readers who wish to measure their texts to the intended readers. The applications and implications of the resulting outcomes from this study can benefit textbook composers, who are working on the new set of K-12 textbooks, such as students, teachers, and publishers.
Keywords: readability assessment, linguistic features, morphological analysis

INTRODUCTION

In the 19th century, various investigations on readability began and their findings contributed to many social domains. In the English-speaking world, readability has been applied in text categorization and evaluation. Various formulae for English text readability have been developed. Although being the 13th spoken language in the world. However, Vietnamese is an under-represented language in the field of readability due to its limitations on corpora and natural language processing. Therefore, building a formula to gauge Vietnamese text readability is a crucial task for those who are working in the field of computational linguistics. Especially, in the current settings while Vietnam is carrying out its education reforms of the curriculum and textbooks from primary to high school levels. The current study aims at forming a formula for Vietnamese text readability.

To date, several formulas have been developed for English text readability. However, those formulas cannot be applied directly into Vietnamese since the difference in language typology. Text readability depends on linguistic factors of written texts, which we consider as linguistic-internal factors.

Our study targets qualitatively and quantitatively developing a readability scale for users/readers who wish to measure their texts to the intended readers. The applications and implications of the resulting outcomes from this study can benefit textbook composers, who are working on the new set of K-12 textbooks, such as students, teachers, and publishers.

METHOD

A Brief Survey on Existing Methods

There is a significant body of research on readability of text that has been developed in the last decades [Kevyn, 2014]. Traditional approaches mainly rely on computing difficulty measures. These measures are normally computed on two main factors, either on *the familiarity of linguistic units* such as words and phrases, or on the *complexity of syntax*. These factors are often combined to devise readability formulas so as to make their application straightforward. The readability scores obtained by these formulas help evaluate the readability or difficulty of traditional texts. The most widely used traditional formula is the Flesch-Kincaid score [Kincaid et al., 1975], which is

*FK_score = 0.39 * (AverageWordPerSentence) + 11.8 * (AverageSyllablePerWord) - 15.59*

This formula is the basis of many similar variants which have been developed over the years. However, as stated above, these formulas are all specific to English and not readily transferable to different languages, especially languages of different types such as the Vietnamese language.

Recently, with the advance of many machine learning methods and the availability of training data, there has been an increasing interest in applying artificial intelligence (AI) based approaches to readability assessment. In these learning-based approaches, there are three main steps. The first step is *corpus acquisition*, where a gold-standard corpus of individual texts is constructed. This corpus is representative of the target language, genre or other aspects of the text that need to be evaluated. The acquired corpus is normally manually annotated by linguistic experts, with the help of computer scientists. It is then divided into a training set and a test set. The training set is used to develop automated machine learning models which are learned from examples. The test set, whose examples are served as unseen samples, is used to evaluate the performance of the learned models. These models sometimes can be tuned on a different test set, which is usually called a validation set or a development set.

The second step is *feature extraction*, sometimes called *featurization*. This step concerns defining and extracting a set of important features that best represent the text under assessment. The feature sets are often proposed by experts with a deep domain knowledge, which are salient to the target readability prediction task. Most of the time, the feature sets are gone through a trial and error process. Note that once a feature set is defined, their feature instances are extracted/computed by an automated software system.

The third step is *model learning*. In this step, a machine learning model is used to learn a mapping from a text to its gold-standard label. This model relies on the features which are extracted from the previous step, both in the training process or inference process. The main assumption of model-based learning is that if the texts are drawn from the same statistical distribution, then if a statistical-based machine learning model performs well on the training data, it will perform well on unseen data too. That is, it helps predict accurately the readability of an unseen text.

Our Proposed Method

In this work, we adopt the machine learning approach to readability assessment, taking into account specific features of the Vietnamese language.

Feature Extraction

In the first step, we design a set of salient features which are suitable for assessing the readability of a text. For each text, we compute the following features:

- The average sentence length in characters
- The average sentence length in words (after performing word segmentation)
- The ratio of concrete/abstract nouns, which is the number of concrete/abstract nouns divided by the number of tokens
- The ratio of proper nouns
- The ratio of adjectives
- The ratio of clauses, which is approximated by the ratio of prepositions
- The ratio of Sino-Vietnamese words; a Sino-Vietnamese word is a word or morpheme of the Vietnamese language borrowed from Chinese
- The ratio of pure old Chinese words, which originated from Chinese
- The ratio of pure Vietnamese words
- The ratio of French loanwords, which are borrowed from French
- The ratio of unknown words, which are not in the standard lexicon
- The identity of most frequent unigrams with a cutoff threshold of 2
- The identity of most frequent bigrams with a cutoff threshold of 2

Note that the last two factors are feature templates, which can generate many feature instances. For example, if we consider a text of 4 words "*difficulty assessment of text*", then the bigram feature templates would generate the following features: "*difficulty assessment*", "*assessment of*", and "*of text*". All n-gram features whose frequency not less than 2 are retained and fed into the machine learning model.

In order to compute these features automatically, we need to develop some core pre-processing modules, including

- A sentence segmentation module which splits a text into multiple sentences
- A word segmentation module which splits a sentence into lexical units (words)

These modules make use of advanced computer algorithms, as described in scientific publications:

- [Le-Hong and Ho, 2008] for automatic sentence segmentation
- [Le-Hong et al., 2008] for word segmentation
- [Le-Hong et al., 2010] for part-of-speech tagging
- [Le-Hong et al., 2017] for clause extraction and tagging

Due to space limitation, we refer the interested reader to the document above for details. In summary, this project result is built upon many essential works that we have performed over the last ten years.

Machine Learning Model

There are a variety of machine learning models for supervised learning which can be used for readability assessment, ranging from linear models to stronger non-linear ones. Given the size and nature of the training corpus, we chose to use a linear classification model namely logistic regression. This model is proven to be both fast and efficient for the problem concerned in this project.

We present briefly the mathematical formulation of this model as follows. Let x be an input text and y be its label. After featurization, the input x is represented by a real-valued vector $f(x)$ of size d, where d is the domain dimension which can be numerous and grows according to the size of the training data. The label y is binary, taking a value

of either 0 (for easy) or 1 (for difficult). In this model, we compute the conditional probability distribution of label *y* given *f(x)* by using the sigmoid function (also called the logistic function), as follows:

*P(y = 1 | f(x); w) := 1 / [1 + exp(-w * f(x))],*

where the parameter vector *w* is also of size *d*, and w * f(x) is the inner product of two vectors *w* and *f(x)*. The right-hand side formula is called the sigmoid function of *w * f(x)*:

sigmoid(u) = 1 / [1 + exp(-u)].

Once the probability that the label is assigned value 1 is computed, we can easily compute the probability that it takes value 0:

*P(y = 0 | f(x); w) = 1 - sigmoid[w * f(x)].*

Given a training dataset which is composed of N training samples {(x_1, y_1), (x_2, y_2),..., (x_N, y_N)}, we can estimate the parameter vector *w* by solving a mathematical optimization model by applying the maximum likelihood principle:

L(w) := P(y_1 | f(x_1); w) P(y_2 | f(x_2); w) ... P(y_N | f(x_N); w) → max

Advanced numerical algorithms such as gradient-based methods or Newton-Raphson methods are shown to be very efficient to solve this optimization problem.

Once the parameter vector *w* is learnt, it can be used to make a prediction for future data samples with a linear time complexity, which is very fast. To assign an unseen text *x* to label 1 or 0, we just need to evaluate *f(x) * w*; if this quantity is greater than a threshold value, for example 0.5, then y is predicted to be 1, otherwise it is predicted to be 0.

In essence, the parameter vector *w* encodes the importance of every salient feature in the prediction model. The greater (in absolute value) a parameter value *w_j* is, the more important the corresponding feature *f_j(x)* is.

Linguistic Resources Construction

In this section, we introduce the construction of two types of resources. The first is lexical resources used for text preprocessing and lexical feature extraction, and the second is the corpora containing the texts to assess the readability.

As mentioned in Section 2.1, the difficulty measures are computed based on two factors, which are the familiarity of linguistic units and the complexity of syntax. In Section 2.2, a list of features is consequently proposed for the construction of a model predicting the difficulty level of a text. For extracting these features, we need the following lexical resources:

- A word list with part-of-speech (POS) information for the tasks of word segmentation and POS tagging;
- From the above list, we can extract a list of concrete/abstract nouns and a list of adjectives;
- A list of conjunctions;
- A list of Sino-Vietnamese words;
- A list of pure old Chinese words;
- A list of pure Vietnamese words;
- A list of French loanwords.

For the Vietnamese word list with POS information, we make use of the Vietnamese Computational Lexicon (VCL) introduced in [Nguyen et al, 2006] and [Vu and Nguyen, 2008]. Each sense of a word entry is associated with several linguistic characteristics: morphological information, word category and

subcategory, subcategorization frames for verbs and semantic descriptions (meaning, semantic constraints, definition and usage examples). Below is an example of the first sense of the word entry "chạy" (to run).

1. chạy (V) [người, động vật] di chuyển thân thể bằng những bước nhanh, mạnh và liên tiếp

Morphological

WordType --> simple word

Syntactic

Category --> V

Subcategory --> Vi

FrameSet --> Sub+V

SyntacticFunction --> Sub

SyntacticConstituent --> NP

Before --> R: đang

Semantic

Logical constraint

CategorialMeaning --> Activity

Semantic constraint

Sub --> Agt{Person, Animal}

Def--> [người, động vật] di chuyển thân thể bằng những bước nhanh, mạnh và liên tiếp

Exa--> *cậu bé đang chạy*

==========

The VCL data is encoded in XML format. This dataset contains about 42,000 entries. From this dataset, we built a tool for extracting all the words found in the studied corpus and their characteristics. The word category and subcategory can be extracted directly from the dataset, while the attribute of a noun being abstract or not is reconstituted from the meaning category in the lexicon and the semantic tree of the lexicon guidelines. However, for several words in VCL, these descriptions are missing. We have filtered these words out to complete our information.

The lists of Sino-Vietnamese words, old Chinese words and French loanwords are built from many Sino-Vietnamese dictionaries and research works. An unexhausted list of works from which the etymological information of the words can be registered as follow:

- Từ điển yếu tố Hán Việt thông dụng (Dictionary of sino-vietnamese everyday usage elements) / chủ biên : Hoàng Văn Hành ; những người biên soạn : Phan Văn Các, [et al.], Hà Nội : Nhà xuất bản khoa học xã hội, 1991.
- 越南语双音节汉越词特点研究 : 与汉语比较 = Yuenanyu shuangyinjie HanYueci tedian yanjiu : yu Hanyu bijiao /罗文青. 罗文青著. ; Wenqing Luo, 世界图书广东出版公司, Guangzhou : Shi jie tu shu Guangdong chu ban gong si, 2011.
- Từ điển từ Hán Việt (2007) của Lại Cao Nguyện (chủ biên) và Phan Văn Các, Nhà xuất bản KHXH.

- Từ điển Các từ tiếng Việt gốc Pháp (Dictionnaire des termes Vietnamiens d'étymologie française) của các tác giả Nguyễn Quảng Tuân và Nguyễn Đức Dân, xuất bản năm 1992.
- Từ gốc Pháp trong tiếng Việt / Vương Toàn, Hà nội : NXB. Khoa học xã hội, 1992.

Concerning the preparation of the corpora for learning and testing models, we have collected the documents from textbooks, then built a tool to format each document in XML. An interface has been defined to annotate the difficulty level of each text.

Software Implementation

In this section, we present the general information about the software system that we have implemented in this project to build an automated system for the readability assessment of text extracted from textbooks.

In order to build a software system that is capable of assessing the readability level of a text efficiently, we need to build from scratch a variety of software modules. These modules can be grouped into 5 main categories as shown in the following table:

Table:

	Category	**Description**	**Modules**
	Core linguistic preprocessing	This category contains modules for segmentation of a text into sentences, and segmentation of sentences into lexical units or words.	- Sentence segmentation - Word segmentation
	Core linguistic processing	This category contains modules for word category tagging (part-of-speech tagging) at the sentence level, and clause segmentation.	- Part-of-speech tagging - Clause segmentation
	Feature extraction	This category contains modules for extracting important linguistic features for readability assessment. There are two types of features, namely discrete features and continuous features.	- Some summary statistics about lengths (average sentence length in words, average sentence length in characters) - Ratio of Sino-Vietnamese words - Ratio of pure Chinese originated words - Ratio of French originated words - Ratio of unknown words - Ratio of common nouns - Ratio of proper nouns - Ratio of adjectives

			- Ratio of prepositions/clauses - Unigram features - Bigram features - Word embeddings
	Model Estimation	This category contains modules for automatic assessment of readability of a text. Two classification models are investigated, namely logistic regression and neural network.	- Training and prediction with logistic regression - Training and prediction with feed-forward neural network model
	Web services	This category contains modules for software integration and demo.	- Data indexing service - Sentence segmentation service - Word segmentation service - Part-of-speech tagging service - Readability assessment service - Demo website (using Java Enterprise technologies)

The underlying assessment model is trained on a set of literature text extracted from the textbook of Grade 4 classes (Level 04).

For example, when a user enters the following text:

Ăng-co Vát là một công trình kiến trúc và điêu khắc tuyệt diệu của nhân dân Cam-pu-chia được xây dựng từ đầu thế kỉ XII. Khu đền chính gồm ba tầng với những ngọn tháp lớn. Muốn thăm hết khu đền chính phải đi qua ba tầng hành lang dài gần 1500 mét và vào thăm 398 gian phòng. Suốt cuộc dạo xem kì thú đó, du khách sẽ cảm thấy như lạc vào thế giới của nghệ thuật chạm khắc và kiến trúc cổ đại. Đây, những cây tháp lớn được dựng bằng đá ong và bọc ngoài bằng đá nhẵn. Đây, những bức tường buồng nhẵn bóng như mặt ghế đá, hoàn toàn được ghép bằng những tảng đá lớn đẽo gọt vuông vức và lựa ghép vào nhau kín khít như xây gạch vữa.

The system gives final and intermediate analysis results, which are shown graphically.

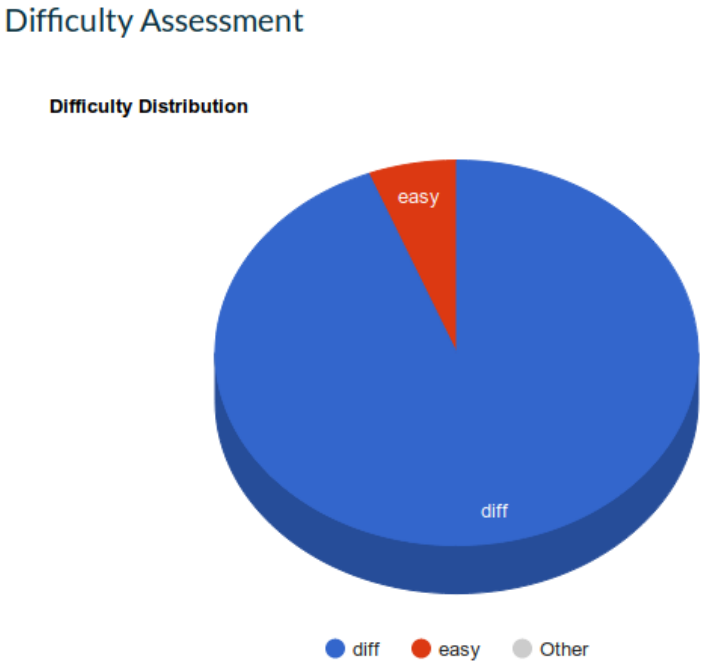

Figure:

A probability distribution of difficulty level is estimated and presented, including two main outcomes: easy or difficult with a proportional ratio. The numbers will be shown when the user hovers the mouse over the corresponding graphical parts.

Some intermediate analyses are also presented for users to check. The first one is sentence analysis, which contains the results of the sentence segmentation module. In the example above, this is a list of six sentences:

1. *Ăng-co Vát là một công trình kiến trúc và điêu khắc tuyệt diệu của nhân dân Cam-pu-chia được xây dựng từ đầu thế kỉ XII.*
2. *Khu đền chính gồm ba tầng với những ngọn tháp lớn.*
3. *Muốn thăm hết khu đền chính phải đi qua ba tầng hành lang dài gần 1500 mét và vào thăm 398 gian phòng.*
4. *Suốt cuộc dạo xem kì thú đó, du khách sẽ cảm thấy như lạc vào thế giới của nghệ thuật chạm khắc và kiến trúc cổ đại.*
5. *Đây, những cây tháp lớn được dựng bằng đá ong và bọc ngoài bằng đá nhẵn.*
6. *Đây, những bức tường buồng nhẵn bóng như mặt ghế đá, hoàn toàn được ghép bằng những tảng đá lớn đẽo gọt vuông vức và lựa ghép vào nhau kín khít như xây gạch vữa.*

The second intermediate analysis is contains part-of-speech tagged text, where each sentence is labeled with words and their corresponding word categories in context:

1. *Ăng-co/Np Vát/Np là/V một/M công_trình/N kiến_trúc/N và/CC điêu_khắc/N tuyệt_diệu/N của/E nhân_dân/N Cam-pu-chia/Np được/R xây_dựng/V từ/E đầu/N thế_kỉ/M XII/Np PUNCT/PUNCT*
2. *Khu/Nc đền/N chính/Np gồm/V ba/M tầng/N với/E những/L ngọn/N tháp/N lớn/A PUNCT/PUNCT*

3. *Muốn/V thăm/V hết/N khu/N đền/N chính/Np phải/V đi/V qua/E ba/M tầng/N hành_lang/N dài/A gần/A 1500/M mét/Nu và/CC vào/E thăm/V 398/M gian/N phòng/N PUNCT/PUNCT*
4. *Suốt/A cuộc/N dạo/V xem/V kì_thú/N đó/P PUNCT/PUNCT du_khách/N sẽ/R cảm_thấy/V như/C lạc/V vào/E thế_giới/N của/E nghệ_thuật/N chạm_khắc/V và/CC kiến_trúc/V cổ_đại/N PUNCT/PUNCT*
5. *Đây/P PUNCT/PUNCT những/L cây/N tháp/N lớn/A được/R dựng/V bằng/E đá_ong/N và/CC bọc/V ngoài/A bằng/E đá/N nhẵn/N PUNCT/PUNCT*
6. *Đây/P PUNCT/PUNCT những/L bức/Nc tường/N buồng/N nhẵn/A bóng/A như/A mặt/N ghế/N đá/N PUNCT/PUNCT hoàn_toàn/A được/V ghép/V bằng/E những/L tảng/N đá/N lớn/A đẽo_gọt/V vuông_vức/N và/CC lựa/V ghép/V vào/E nhau/N kín/A khít/N như/C xây/V gạch/N vữa/N PUNCT/PUNCT*

It can be shown in the above result, each token is labeled with a tag designating a part-of-speech category, for example *N* is a common noun, *Np* is a proper noun, *A* is an adjective, *V* is a verb, *E* is a preposition, and so on. To develop the part-of-speech tagging module, we use a linguistic corpus of more than 10,000 sentences which are manually word segmented and tagged by linguists at the Vietnam Center of Lexicography (Vietlex). This corpus is a result of the VLSP 2010 project, funded by the state whose objective is to build fundamental resources and tools for processing Vietnamese text and speech.

The third intermediate analysis contains some linguistic features which are essential for readability assessment, as shown in the following figure:

Feature Analysis

Feature	Value
F00_TOKENS_PER_SENTENCE	22.833
F01_CHARACTERS_PER_SENTENCE	81.667
F02_RATIO_OF_COMMON_NOUNS	0.331
F03_RATIO_OF_PROPER_NOUNS	0.023
F04_RATIO_OF_ADJECTIVES	0.09
F05_RATIO_OF_PREPOSITIONS	0.083
F06_RATIO_OF_SINO_VIETNAMESE	0.263
F07_RATIO_OF_CHINESE	0.023
F08_RATIO_OF_FRENCH	0.038

Figure:

Here, some ratios of sentence lengths in tokens and in characters as well as some etymological features and syntactic features are shown.

RESULTS

The automatic text readability assessment is performed in English, German, Swedish, Japanese and Chinese. In contrast, research on readability of Vietnamese text is quite limited. The purpose of this study is to systematically analyze the impact of

linguistic features for assessing the readability level of Vietnamese texts for K-12 learners. More specifically, we designed various features at different levels: morphology, part-of-speech, syntactic, and discourse, and applied classification models for potentially predicting the reading levels of Vietnamese textbooks for elementary, junior high, and senior high school students. In the current model, we have tested on selected linguistic features (in black) at different levels, as can be seen in Table # below. We further regressed these features for different readability levels and selected significant features.

Table : Linguistic features included in the model

Level	**Domain**	**Feature**
Morphology	Word complexity	- The ratio of Sino-Vietnamese words; a Sino-Vietnamese word is a word or morpheme of the Vietnamese language borrowed from Chinese - The ratio of pure old Chinese words, which originated from Chinese - The ratio of pure Vietnamese words - The ratio of French loanwords, which are borrowed from French - The ratio of unknown words, which are not in the standard lexicon - The identity of most frequent uni-grams with a cutoff threshold of 2 - The identity of most frequent bigrams with a cutoff threshold of 2 - Average number of syllables per word per document - Average number of syllables per unique word per document - Percentage of more than two syllable words per document
POS		- The ratio of common nouns, which is the number of common nouns divided by the number of tokens - The ratio of proper nouns - The ratio of adjectives per document - Percentage of unique functional words per document - Number of unique functional words per document - Average number of unique nouns per sentence
Syntactic	Sentence complexity	- The average sentence length in characters - The average sentence length in words (after performing word segmentation) - The ratio of clauses, which is approximated by the ratio of prepositions - Average number of multi-syllable words per sentence - Average length of prepositional phrases per document

Syntactic	Document complexity	- Number of syllables per document - Number of syllable (including punctuations, numerical, and symbols) per document - Percentage of unique nouns per document
Discourse	Entity density Cohesion	- Average number of unique entities per sentence - Number of unique conjunctions per document - Percentage of unique conjunctions per document - Average number of conjunctions per sentence

Based on pilot data from extracted texts from textbooks for grade 4, we fitted the data into the learning model. The T-test p values show that the model achieves high accuracies for level 4.

The sample data for the learning model are 80 texts classified by tertiary linguistic and literature students on a 7-level Likert scale. We asked the students to read those texts and classify them into different levels. The students are also asked to spell out some of the reasons that make the text difficult.

We use the mentioned features to perform regression on various level text data. We select a subset of them at 96% confidence level and derive a readability formula as presented above.

In the next stage of the project, we collected and processed a large body of texts in 17 subjects and evaluated our proposed method of difficulty assessment on these texts. For each subject, we take 80% of texts for training and 20% of texts for testing. Each text is classified into either 'easy' or 'difficult' level. In total, there are 4,930 texts which are processed.

The statistics of subjects, their corresponding number of lessons are given in the following table.

Subject	**Grade**	**Number of texts (lessons)**	**Test Accuracy**	**Test F-measure**
Arts	L04	35	85.71%	85.71%
	L05	35	57.14%	62.85%
Biology	L06	54	83.33%	82.85%
	L07	67	80.00%	78.85%
	L08	64	73.33%	71.85%
	L09	62	100%	100%
	L10	33	71.43%	72.62%

	L11	50	63.64%	63.03%
	L12	56	91.67%	91.48%
Chemistry	L08	13	0%	0%
	L09	54	75.00%	75.52%
	L10	39	87.50%	87.30%
	L11	47	72.72%	72.72%
	L12	40	87.50%	87.30%
National Defense	L10	7	100%	100%
	L11	9	100%	100%
	L12	7	0%	0%
Civic Education	L06	18	100%	100%
	L07	18	100%	100%
	L08	22	66.66%	53.33%
	L09	15	0%	0%
	L10	15	100%	100%
	L11	15	50.00%	33.33%
	L12	10	100%	100%
Geography	L06	28	80.00%	80.00%
	L07	63	73.33%	71.85%
	L08	44	80.00%	80.00%

	L09	45	90.00%	90.32%
	L10	45	90.00%	89.89%
	L11	37	62.50%	64.28%
	L12	43	77.77%	77.22%
History	L06	29	100%	100%
	L07	55	75.00%	74.47%
	L08	33	57.14%	59.04%
	L09	35	100%	100%
	L10	40	87.50%	87.30%
	L11	26	75.00%	76.66%
	L12	28	80.00%	80.00%
History - Geology	L04	64	93.33%	93.20%
	L05	53	100%	100%
Informatics	L03	23	100%	100%
	L04	48	100%	100%
	L05	28	100%	100%
	L06	21	100%	100%
	L07	11	0%	0%
	L08	15	100%	100%
	L09	23	75.00%	76.66%

	L10	25	75.00%	76.66%
	L11	22	66.66%	66.66%
	L12	22	100%	100%
Literature	L06	36	71.42%	83.33%
	L07	33	71.42%	71.42%
	L08	33	71.42%	72.62%
	L09	41	87.50%	86.82%
	L10	35	71.42%	71.42%
	L11	37	75.00%	75.00%
	L12	32	100%	100%
Morality	L04	14	0%	0%
	L05	13	0%	0%
Natural - Social Science	L01	35	100%	100%
	L02	34	100%	100%
	L03	64	100%	100%
Physics	L06	27	40.00%	40.00%
	L07	29	60.00%	63.33%
	L08	26	100%	100%
	L09	38	71.42%	72.62%
	L10	42	62.50%	64.28%

	L11	50	100%	100%
	L12	53	75.00%	76.67%
Science	L04	65	86.67%	86.00%
	L05	57	92.30%	92.106%
Technology	L06	26	100%	100%
	L07	57	84.61%	84.61%
	L08	57	100%	100%
	L09	6	100%	100%
	L10	58	76.92%	76.08%
	L11	33	85.71%	86.34%
	L12	30	60.00%	63.33%
Vietnamese	L01	159	75.86%	75.68%
	L02	270	86.53%	86.60%
	L03	257	65.30%	65.30%
	L04	288	77.35%	77.32%
	L05	266	59.61%	59.75%

CONCLUSION

In this study, we have systematically considered the impact of linguistic features on readability for Vietnamese texts. We used a set of selected texts from the primary school, junior high school and high school Vietnamese textbooks as our input data set. The set of linguistic features accounted for various aspects, such as morphological features, POS features, syntactic features and discourse features.

The current pilot study does not allow us an insight into the best model yet. However, in the next phase of the research, we suggest that we will fit all the annotated texts from all the textbooks from grade 1 to grade 12 into the learning model. That way, we could achieve a highly predictive and accurate model for the majority of the Vietnamese text

readability for generally educational purposes. We also expect to extend this research for assessing a wide range of Vietnamese texts in other domains with various implications.

REFERENCES

[Kevyn, 2014], Kevyn Collins-Thompson, "*Computational Assessment of Text Readability: A Survey of Current and Future Research"*, Working Draft, 2014.

[Kincaid et al., 1975], J. Peter Kincaid Robert P. Fishburne Jr. Richard L. Rogers Brad S. Chissom, "*Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel"*, Research report, University of Central Florida, 1975.

[Le-Hong et al., 2008] Le-Hong, P., T M H. Nguyen, A. Roussanaly, and T V. Ho, "*A hybrid approach to word segmentation of Vietnamese texts*", Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Tarragona, Spain, Springer, LNCS 5196, pp. 240--249, 2008.

[Le-Hong and Ho, 2008] Le-Hong P. and Tuong-Vinh Ho, "*A maximum entropy approach to sentence boundary detection of Vietnamese texts*", Proceedings of RIVF 2008, IEEE, Ho Chi Minh City, Vietnam, 2018.

[Le-Hong et al., 2010] Le-Hong, P., T M H. Nguyen, M. Rossignol, and A. Roussanaly, "*An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts*", Actes du Traitement Automatique des Langues Naturelles (TALN-2010), Montreal, Canada, 2010.

[Le-Hong et al., 2017] P. Le-Hong, T-H Pham, X-K Pham, T-M-H Nguyen, T-L Nguyen, M-H Nguyen, "*Vietnamese Semantic Role Labeling*", VNU Journal of Science: Computer Science and Communication Engineering, Vol. 33, No. 2, pp. 1-21, 2017.

[Nguyen et al., 2006] Nguyễn T. M. H., Vũ X. L., Romary L., Rossignol M., "A Lexicon for Vietnamese Language Processing", Language Resources and Evaluation, Special Issue: Asian Language Processing: State-of-the-Art Resources and Processing, Springer Netherland, vol. 40, no. 3-4, p. 291-309, 2006.

[Vu and Nguyen, 2008] Vũ X. Lương, Nguyễn T. M. Huyền, "Building a Vietnamese Computational Lexicon", Proceedings of the National Symposium on Research, Development and Application of Information and Communication Technology, Vietnam, 2008.