AFFECTION OF THE PART OF SPEECH ELEMENTS IN VIETNAMESE TEXT READABILITY

Điệp Thi Nhu NGUYỄN
Vietnam National University Ho Chi Minh City, Vietnam
nhudiep2004@gmail.com

An-Vinh LƯƠNG
Vietnam National University Ho Chi Minh City, Vietnam
anvinhluong@gmail.com

Điền ĐINH
Vietnam National University Ho Chi Minh City, Vietnam
ddien@fit.hcmus.edu.vn

Abstract

While English text readability has been studied for a long time, investigating text readability in Vietnamese, a low-resourced language with poorresearch technologies and data sets questionable of international importance, is at its beginnings. In readability research, it is generally the "word" that has been carefully investigated. Based on the comparison of elements affecting readability of the "word" unit in English, we determine the parts of speech (POS) in Vietnamese that were found to influence Vietnamese text readability. In this study, prose texts in Vietnamese textbooks at different difficulty level were taken as the data to find out the POS frequencies and their correlations. In terms of frequency, our findings can initially assist users when editing documents, reforming textbooks, and question banks for native Vietnamese in general and foreigners in particular. Even more important, with these findings we can identify those linguistic elements that are considered the "potential" POS affecting Vietnamese text readability, and make grounds for further studies.

Keywords: text readability; parts of speech; Vietnamese textbooks; elementary level

Povzetek

Medtem ko je že precej vemo o bralni pismenosti angleških tekstov, pa so takšne raziskave na tekstih v vietnamščini šele na začetku. Večina raziskav o bralni pismenosti se osredotoča na "besedo". Na osnovi primerjav elementov, ki vplivajo na bralno pismenost na nivoju besede v angleščini, smo v naši raziskavi določili besedne vrste

Acta Linguistica Asiatica, 9(1), 2019.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.9.1.105-118

(angl. “parts of speech”, POS), pri katerih smo zaznali, da vplivajo na bralno pismenost v vietnamščini. V raziskavi so bili obravnavani učbeniki vietnamščine in sicer njihovi prozni teksti, iz katerih smo ocenili pojavnost posameznih besednih vrst in njihovo korelacijo z različnimi težavnostnimi nivoji. Že same informacije o pojavnosti lahko pripomorejo k boljšemu razumevanju bralne pismenosti in so v pomoč pri pripravi in urejanju dokumentov, pisanju učbenikov, sestavljanju vprašalnikov tako za domače govorce, še posebej pa za tuje govorce vietnamščine. Še bolj pomembni pa so seveda pridobljeni podatki o jezikovnih elementih, ki so označeni kot besedne vrste, ki potencialno vplivajo na bralno pismenost v vietnamščini. Slednji predstavljajo osnovo za vse nadaljne raziskave.

Ključne besede: bralna pismenost; besedne vrste; učbeniki vietnamščine; začetna stopnja

1 Introduction

The studies of readability have been done since the early nineteenth century. Among these achievements are the formulas for measuring readability, which are used as a tool for determining the complexity of the text. Therefore, they can help users select an appropriate text with different reading levels for the readers in efficiently, saving time and labor. The results of the research have applied in various areas of society, such as the integratedly measuring the Flesch formula in Microsoft Office software or the same with the formulae: Flesch-Kincaid, Cohmetrix, Idicies, Lexile Measures, etc. in the Common European Framework of Reference.

In forming a formula or a tool to measure text readability, linguistic elements or linguistic components in a particular text play a very important role, as shown in a lot of readability research, such as Gray and Leary (1935), Lorge (1939), Rudolf Flesch (1943; 1946; 1948), Graesser et al. (2004), and McNamara et al. (2014). These linguistic elements were gained through analyses on the shallow/surface features on one hand, such as the average length of words by the number of syllables, the average numbers in a sentence, or the frequency of words; and the deep features of the language on the other hand, such as the parsed syntactic features, the language modeling features, or the part of speech- based features.

With the scope of this article, we first define the part of speech as the linguistic elements affecting the text readability in Vietnamese based on the contrast of the linguistic elements affecting text readability in English, we survey and evaluate readability influences of part of speech (POS) elements of prose texts in Vietnamese subject textbooks for elementary school-aged children based on the several statistic measures.

Results of this study are expected to be useful to writers, editors, and especially to teachers and learners of Vietnamese, who compile or select lectures and banks of questions based on the grade level.

2 Methodology and corpus

Our corpus represents prose texts in Vietnamese textbooks for elementary school children (grades 2–5) that were published by Education Publisher in 2016. In the preprocessing, we have decided to leave out the texts that were in forms of questions, puzzles or drawing annotations, and therefore were left with 209 texts in the end. Those texts are all estimated to provide children with general knowledge and help them practice reading skills. Linguistic elements with surface features are described in Table 1 below:

Table 1: Vietnamese textbook corpus

Grade	Number of Texts	Number of Words	Number of Sentences
2	67	57 – 251	5 - 40
3	62	112 - 279	8 - 35
4	40	144 - 520	7 - 47
5	40	111 - 381	4 - 52

We used the "CLC-Vietnamese-Toolkit"[1], generated by Computational Linguistics Center, University of Science, HCMC, to handle the POS in each text, and calculate their frequency. Besides, the relationship between the POS with the text readability was also investigated.

3 Affection of linguistic elements in text readability

3.1 Linguistic elements affecting text readability in English

Gray and Leary's (1935) identified 288 elements affecting English text readability, and these elements were classified into four main categories: (I) format or mechanical features, (II) general features of organization, (III) style of expression and presentation, and (IV) content (Gray & Leary, 1935).

Within the scope of their study, they have identified 82 language elements that function as the "potential elements" affecting text readability by investigating the linguistic elements of style of expression and presentation alone. These elements are classified under three different units, namely word, sentence and paragraph/passage.

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

Among them, 41 elements affecting text readability at word level were counted. With the aim to conduct an experimental research based on quantitative enumeration, 14 out of those 41 language elements were left out of further analysis due to the following reasons: (i) the linguistic elements do not meet the experimental process; (ii) they has not been formed by the clear definitions yet, and (iii) these linguistic elements cannot be measured or counted objectively in largely analyzed cases from the corpus.

Based on this elementary work, many studies have investigated and developed the language elements affecting English text readability. Examining the same "structural elements" as Gray and Leary (1935), Lorge (1939) added an additional variable, "a weighted index of word difficulty". Lorge believed that prepositions played important role to measure syntactic complexity in English. He suggested the readability formula which adjusts weights and uses various combinations of two variables such as (i) prepositional phrases and different hard words, (ii) average sentence length and different hard words, and (iii) the number of prepositional phrases and average sentence length (Lorge, 1939).

In creating a regression formula that could with some accuracy distinguish levels of difficulty for both children's and adults' reading material, besides sentence length Rudolf Flesch (1943) added two other variables: the number of affixes and a variable used in Gray and Leary. The number of personal pronouns, which Flesch limited to gendered (non-neutral) pronouns, were represented by the human interest factor of the texts (Flesch, 1943). Flesch (1948) defined the idea of personal words somewhat differently in order to codify human interest: "All nouns with natural gender; all pronouns except neuter pronouns; and the words people (used with the plural verb) and folks". To this, Flesch added another factor, which he called "personal sentences". This factor was intended to be a measure of the "conversational quality and the story interest" of the passage analyzed (Flesch, 1948). The Art of Readable Writing (Flesch, 1949) was a popular success as a "how-to" book about writing, successful enough that a quarter of a century later the book was reissued in a new, expanded edition (Flesch, 1974). The Reading Ease formula was adapted for use by the United States Military using the same factors but somewhat different weights (Kincaid et al., 1975) and can be found to this day as a tool in the most popular word processing program in the world, Microsoft Word.

Coh-Metrix is a major departure from both the classic formulas and cloze. It is a computational tool that facilitates the formulation and testing of hypotheses about readability and other reading comprehension issues: "Coh-Metrix ... analyzes texts on over 200 measures of cohesion, language, and readability. Its modules use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components that are widely used in computational linguistics" (Graesser et al., 2004). In classifying part-of-speech, McNamara et al. (2014) presented that Coh-Metrix permits more sophisticated measures of grammatical complexity, it

can count the mean number of modifiers in noun phrases and the mean number of words that occur before the main verb. In particular, Coh-Metrix includes indices for various linguistic features that can be considered markers of cohesion, for example, it contains an index for measuring the number of causal connectives- connectives indicating the logical relations between parts of the text (e.g., because, so). It also contains an index relating causal particles (e.g., due to, therefore, if) to causal verbs. The hypothesis is that the higher the ratio of causal particles to causal verbs, the more cohesive a text is, since it suggests that there are more explicit indications of how events and actions are interrelated (McNamara, Graesser, McCarthy, & Cai, 2014, pp. 62-68).

Thus, language elements in general, and the parts of speech in particular, have been investigated more and more deeply in English text readability to meet the practical needs. However, it is important to note that there are many differences between English and Vietnamese, ranging from morphological typology (morphemes, word boundaries, the word forms, for example “anh” in Vietnamese means “elder brother” in English), and sentence structure (theme-rheme relationship), to the differences in phonetics and phonology. Therefore, adjustments to the existing model should be made, and comparisons and contrasts between these two languages are crucial in this case (Đinh, 2006). Hence, by comparing the similarities and differences of the linguistic elements between Vietnamese and English in the word unit, this article selects and surveys the POS elements at the word unit from the above-mentioned corpus.

3.2 Linguistic elements affecting text readability in Vietnamese

3.2.1 Lexico - grammatical category

Language vocabularies are generally very large and it is thus reasonable to further divide words into subclasses to make the word-formation rules and those of their usage more comprehensible. There are several ways to do so. For example, words can be further divided in terms of (1) their meanings; namely some words convey one meaning while others are polysemantic, in terms of (2) their origin, where they can be classified into cognates and borrowed words, (3) according to the frequency of their usage, where common, everyday words are used more often than words of slang, dialectal expressions, technical terms, and others. Words can also be divided (4) based on their word-forms into monosyllabic and polysyllabic words, or else into single and compound words, and nonetheless (5) according to their first letter, as in dictionaries.

In Vietnamese, however, there is another crucial way of word classification, which is based on words’ lexical meanings together with their grammatical functions. It is called lexico-grammatical category (Nguyễn, Đoàn, & Nguyễn, 2008, p. 242).

Each grammatical category includes a set of different forms of a word, but each lexico-grammatical category includes a set of words. The process of determining grammatical category generally begins with considering possible forms of a word to determine their number; for example, in English, book (singular) with books (plural). Only then are words categorized into content words (nouns, verbs, adjectives, adverbs…) and function words (articles, prepositions, conjunctions…). On the other hand, applying lexical-grammatical category means that a word carries a unified form and is as such classified based on its general meaning and grammatical characteristics. Following this, Vietnamese word are divided into either "lexical words" or "form words", with the two categories being comparable to content words and function words respectively.

To avoid the confusion on the classification criteria, we have decided to analyze our corpus and determine POS elements based on lexical- grammatical category, and following the POS classification conducted by the Committee of Social Science (1993). According to the Committee of Social Science, "parts of speech include words with the same general meaning and grammatical characteristics […] The general meaning of Vietnamese words are reflected in their grammatical characteristics. However, their characteristics, in such an isolating language like Vietnamese, are not shown in the phonology but their collocations with other words" (Vietnam Committee of Social Science, 1993, p. 66).

In this classification, lexical words convey the "real meaning" or the "lexical meaning" of objects, and point at the phenomena which establishes the connection between words and objects. In terms of grammar, lexical words can work as "theme" or "rheme" in a sentence. With two lexical words, it is absolutely possible to make a simple sentence.

(1) **Xe** *chạy*
'**Cars** *are moving.*'

(2) **Lúa** *tốt.*
'**The rice** *is growing well*'

On the other hand, form words in Vietnamese do not convey any real meanings, and do not connect to any objects or phenomena. These words themselves cannot function as main parts of a sentence, but have to go with lexical words to make a sentence; hence, they convey grammatical meaning such as time (example (3)) or degree (example (4)).

(3) Xe ***đã*** chạy.
'Cars ***have*** gone.'

(4) Lúa ***rất*** tốt.
'The rice is growing ***very*** well.'

Furthermore, form words can carry additional meanings.

(5) Lúa mùa ***và*** lúa chiêm đều rất tốt.
'The winter rice ***and*** the summer rice grew very well.'

(6) Lúa ***của*** hợp tác xã đó tốt.
'The rice ***of*** that cooperative grew well.'

In order to make the classification more effective and useful in forming sentences, lexical words and form words are divided further into two groups. Lexical words are categorized into nouns, verbs and adjectives; whereas form words are classified into adjuncts and conjunctions. In addition to these categories, we also make the use of pronouns, while modifiers, and interjections are the two categories that belong to both lexical words or form words, and differ from the category of pronouns. To sum up, part of speech in Vietnamese are categorized into eight main groups, of which former six groups are subdivided as follows [2]:

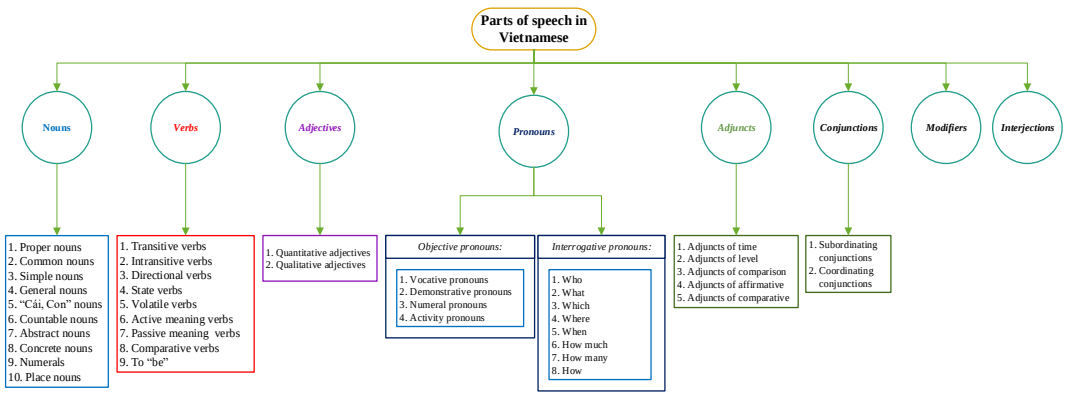

Figure 1: Parts of speech in Vietnamese

3.2.2 POS elements affecting text readability in Vietnamese

Word class is a hierarchical system in which a category consists of smaller categories. Vietnamese words can be divided into the two main categories, cf. lexical words and form words. Each category can be devided further based on the parts of speech. This significance covers a narrower scope of a word, but the meaning remains the general syntactic meaning (Mostafa & Pooneh, 2012, p. 270). Lijun, Martin, Matt and Noemie (2010, re-extracted from Heliman et al. (2007) and Leory et al. (2008)) show that the characteristics of the part of speech in a text prove very useful in determining text readability.

[2] Categorized according to Vietnamese Grammar (1993, pp. 67–95).

To determine the influences of the POS on readability in Vietnamese texts, we used the automatic supporting tool called CLC-Vietnamese-Toolkit, through which we identified 25 common parts of speech in Vietnamese. There were few cases where identification was impossible, and such words were labeled X (unidentified POS - Unknown). We could then investigate the relationship among different parts of speech in text readability, and labelled them with different grade levels (from grade 2 to grade 5).

The corpus was analyzed to determine the frequency of the parts of speech used in each text of each grade. The data showed that some parts of speech were not used at all, and hence the lowest frequency is recorded is zero (0). For example, examining 67 texts in grade 2, we found out that proper noun was not used in 20 of the 67 texts, and 22 times is the largest frequency with which this part of speech was used in texts. Therefore the frequency of proper nouns in grade 2 ranges from from the lowest (0) to the highest (22) as we can see from the extracted data in the Table 2 below:

Table 2: The extracted data of parts of speech in Vietnamese primary textbooks

	A	B	C	D	E	F	G	H	I
1	File	Lớp	Pd	Nu	An	PU	ID	Nc	FW
2	2\Tap 1\0 - Ban cho.txt.ws.pos	2	2	3	0	30	0	6	0
3	2\Tap 1\0 - Ban tay diu dang.txt.ws.pos	2	0	0	0	28	0	10	0
4	2\Tap 1\0 - Be Hoa.txt.ws.pos	2	2	0	0	30	0	9	1
5	2\Tap 1\0 - Can voi.txt.ws.pos	2	2	1	0	9	0	3	0
6	2\Tap 1\0 - Cay xoai cua ong em.txt.ws.pos	2	2	0	0	21	0	0	0
7	2\Tap 1\0 - Di cho.txt.ws.pos	2	6	4	0	22	1	0	0
8	2\Tap 1\0 - Dien thoai.txt.ws.pos	2	2	0	1	32	0	4	2
9	2\Tap 1\0 - Doi ban.txt.ws.pos	2	1	0	0	18	0	1	0
10	2\Tap 1\0 - Doi giay.txt.ws.pos	2	2	0	0	21	0	0	1
11	2\Tap 1\0 - Ga ti te voi ga.txt.ws.pos	2	1	0	0	47	0	0	3
12	2\Tap 1\0 - Ha mieng cho sung.txt.ws.pos	2	1	0	0	19	0	0	0
13	2\Tap 1\0 - Lam viec that la vui.txt.ws.pos	2	1	0	0	29	0	0	1
14	2\Tap 1\0 - Mit lam tho.txt.ws.pos	2	6	1	1	57	0	16	1
15	2\Tap 1\0 - Mua kinh.txt.ws.pos	2	1	0	0	20	0	0	0
16	2\Tap 1\0 - Ngoi truong moi.txt.ws.pos	2	0	0	0	17	0	0	0
17	2\Tap 1\0 - Qua cua bo.txt.ws.pos	2	2	0	0	23	0	0	4
18	2\Tap 1\0 - Them sung cho ngua.txt.ws.pos	2	3	0	0	36	0	2	2
19	2\Tap 1\0 - Tren chiec be.txt.ws.pos	2	0	1	0	20	0	0	0
20	2\Tap 1\1 - Co cung mai sat co ngay nen kim.txt.w	2	1	0	0	27	0	0	0
21	2\Tap 1\10 - Ba chau.txt.ws.pos	2	3	1	0	42	0	0	2
22	2\Tap 1\11 - Su tich cay vu sua.txt.ws.pos	2	2	0	0	42	0	1	2
23	2\Tap 1\12 - Bong hoa Niem Vui.txt.ws.pos	2	2	0	0	29	0	10	1
24	2\Tap 1\13 - Cau chuyen bo dua.txt.ws.pos	2	2	0	0	38	0	0	1
25	2\Tap 1\14 - Hai anh em.txt.ws.pos	2	6	0	0	30	0	0	0
26	2\Tap 1\15 - Con cho nha hang xom.txt.ws.pos	2	2	0	0	39	0	22	1
27	2\Tap 1\16 - Tim ngoc.txt.ws.pos	2	4	0	0	42	0	1	0
28	2\Tap 1\2 - Phan thuong.txt.ws.pos	2	3	0	0	31	0	4	0
29	2\Tap 1\3 - Ban cua Noi Nho.txt.ws.pos	2	3	0	0	29	0	10	2
30	2\Tap 1\4 - Bim toc duoi sam.txt.ws.pos	2	2	0	0	42	0	13	1

Using the CLC-Vietnamese-Toolkit, we examine the parts of speech of the texts in each grade. Based on the statistics, their frequency was calculated, and results are listed in Table 3:

Table 3: The frequency of the POS elements affecting text readability in Vietnamese - prose corpus, primary textbooks

No.	Part of speech	POS	Grade 2	Grade 3	Grade 4	Grade 5	Total
1	Proper Nouns	Nr	0–22	0–20	0–24	0–23	0–24
2	Countable Nouns	Nc	0–16	1–19	0–34	1–15	0–34
3	Concrete Nouns	Nu	0–4	0–10	0–8	0–4	0–10
4	Temporal Nouns	Nt	0–19	0–22	0–19	0–14	0–22
5	Numerals	Nq	1–18	2–25	3–29	4–24	1–29
6	Common Nouns	Nn	11–83	25–89	38–146	28–107	11–83
7	Directional Verbs	Vd	0–6	0–10	0–8	0–5	0–10
8	State Verbs	Ve	0–11	0–8	0–8	0–10	0–11
9	Comparative Verbs	Vc	0–8	0–6	0–6	0–7	0–8
10	Volatile Verbs	Vv	12–74	17–72	17–104	19–90	12–104
11	Directions	D	0–8	0–11	0–9	0–10	0–11
12	Quantity Adjectives	An	0–2	0–5	0–8	0–4	0–8
13	Quality Adjectives	Aa	1–24	4–33	8–43	6–39	1–43
14	Demonstrative Pronouns	Pd	0–8	0–7	0–12	0–11	0–12
15	Personal Pronouns	Pp	0–23	0–38	0–39	0–33	0–39
16	Adverbs	R	1–31	1–33	3–51	2–45	1–51
17	Prepositions	Cm	1–18	1–29	1–29	3–19	1–29
18	Parallel Conjunctions	Cp	0–17	1–17	4–33	3–22	0–33
19	Subordinating Conjunctions	Cs	0–4	0–4	0–3	0–8	0–8
20	Modifiers	M	0–10	0–9	0–12	0–6	0–12
21	Emotion Words	E	0–4	0–4	0–3	0–2	0–4
22	Foreign Words	FW	0–5	0–7	0–6	0–6	0–7
23	Onomatopoeia	ON	0–0	0–2	0–0	0–0	0–2
24	Idioms	ID	0–1	0–1	0–2	0–1	0–2
25	Unidentified POS	X	0	0	0	0	0

According to the corpus analysis outlined above, we can first quantitatively identify 25 elements which affect text readability. The frequency of each element for each grade (from 2 to 5) and the elementary level are identified. For example, the frequency of “proper nouns” in a text of elementary level, from grade 2 to 5, is from 0 to 24, more specifically, in Grade 2, the frequency is from 0 to 22, 0 to 20 for Grade 3, 0 to 24 for Grade 4, and 0 to 23 for Grade 5. The frequency of elements from other categories can also be identified in a similar way. In each grade, 25 parts of speech can be determined in their scopes, out of which we can see differences among language elements per grade as well as per all grades together.

It can be seen from Table 3 that no text at elementary level are uses unidentified POS, and hence investigating this linguistic element in Vietnamese texts at

intermediate and advanced levels is necessary for robust conclusions. The rest 24 elements from the remaining types can be classified into 3 groups: the frequency at low levels (0-34), the frequency of average (35-68) and the group with a high level of frequency (68-104). This is shown in Table 4 below:

Table 4: POS Elements affecting text readability in Vietnamese – Elementary level

POS Elements affecting text readability in Vietnamese - elementary level		
Elements with low frequency	Elements with average frequency	Elements with high frequency
Proper Nouns	Qualitative Adjectives	Common Nouns
Countable Nouns	Personal Pronouns	Volatile Verbs
Concrete Nouns	Adverbs	
Temporal Nouns		
Numerals		
Directional Verbs		
State Verbs		
Comparative Verbs		
Directions		
Quantitative Adjectives		
Demonstrative Pronouns		
Prepositions		
Subordinating Conjunctions		
Parallel Conjunctions		
Modifiers		
Emotion Words		
Foreign Words		
Onomatopoeia		
Idioms		

Besides investigating the POS frequency, we also examine the correlation of these elements to determine their influences on Vietnamese text readability. We used Pearson Correlation to compute these numbers.[3] In this way, we examined linear relations between the POS elements (independent variables) and Vietnamese text readability (dependent variable) by Pearson correlation coefficient (depicted by r). The value of the correlation coefficient ranges from -1 to 1, with r = 0 (or close to 0) suggesting that there is no or very weak relation between a POS element (x) and Vietnamese text readability (y). In cases when correlation coefficient ranges below 0 ($r < 0$), the two correlate inversely, namely that x increases with the decrease of y and

[3] http://phantichspss.com/he-so-tuong-quan-pearson-cach-thao-tac-phan-tich-tuong-quan-trong-spss.html

the other way around. And finally, in cases when correlation coefficient ranges above 0 ($r < 0$), the two correlates show direct relation; when x increases, y will increase. The correlation analysis results are presented in Table 5.

Table 5: The Pearson correlation between the POS elements and Text readability

Part of speech	*r*	Part of speech	*r*
Demonstrative Pronouns	0.160	**Emotion Words**	**-0.111**
Concrete Nouns	0.167	Countable Nouns	0.206
Quantity Adjectives	0.098	**Common Nouns**	**0.511**
Idioms	0.071	Quality Adjectives	0.443
Proper Nouns	0.232	Numerals	0.355
Foreign Words	0.142	Personal Pronouns	0.017
Directional Verbs	0.052	Adverbs	0.231
Volatile Verbs	0.351	**Onomatopoeia**	**-0.026**
Comparative Verbs	0.255	Modifiers	0.019
Prepositions	**0.509**	Coordinating Conjunctions	0.402
Directionals	0.102	State Verbs	0.207
Temporal Nouns	0.229	Subordinating Conjunctions	0.115

The correlation analysis results show that most of them are positively related; and there are only two negative correlation coefficients with text readability: emotion words (-0.111) and onomatopoeia (-0.026); but the influence of two elements on text readability is relatively low (nearly no affection). Among 22 POS elements with positive correlation coefficients, frequencies of common nouns and prepositions have strong connection with the text readability (0.511 and 0.509). This in other words means that in case of common nouns, about 30% of the change of text readability links to the change in frequency of other nouns in the texts. Similarly, the correlation coefficient of prepositions means that, with all the elements being analyzed, about 26% of the change of text readability is related to the change of the frequency of prepositions.

From the above results we can suggest that the two POS elements, namely prepositions and common nouns are the most influential linguistic elements in Vietnamese text readability, such as polysemantic common nouns or prepositions in ambiguity. AS such they are expected to gain attention in further studies.

4 Comments and conclusion

The survey about the extent to which 25 POS elements affect text readability in prose texts in Vietnamese textbooks for primary pupils at elementary level (easy) can help teachers, editors, and learners to determine the level of difficulty qualitatively. The findings, in this level, show that common nouns and volatile verbs are the elements

with the highest frequency, three of the parts of speech with the medium frequency are qualitative adjectives, personal pronouns, and adverbs. Except for the unidentified POS, the rest of the parts of speech - 19 categories- are used with low frequency. In addition, the correlation coefficient also shows that conjunctions and common nouns are the potential language elements affecting Vietnamese text readability, and their meaning and grammatical structure should be investigated further.

The most important thing in evaluating POS elements affecting text readability is that the corpus must be classified in different levels. However, at present, there is no tool or formula reliable or effective enough to measure the text readability for Vietnamese texts. Therefore, choosing a corpus collected from the textbooks which were already classified into different grade levels for elementary school- aged children is ideal for this study. Besides, there are still some issues in the corpus itself. Although the texts hierarchically divided in increasing levels from Grade 2 to Grade 5, there is no clear distinction. For example, the frequency of temporal nouns in grade 2 and 4 is equal (0 -14), while grade 3 has the highest frequency (0-22) and 5th grade, despite being the highest grade, has the lowest frequency (0-14). Therefore, further studies with a larger corpus for this level as well as in intermediate and advanced levels are necessary.

Text readability in English has been studied since the early 19th century, but investigating text readability in Vietnamese is still the beginning. Therefore, in the future, we will build a larger corpus from multiple materials as well as divide the corpus using both quantitative and qualitative methods to calculate at three levels: basic; intermediate; and advanced. We will also investigate Vietnamese readability more deeply with other linguistic elements. Since then, the analysis of the corpus is more reliable and convincing. It can help the computational linguistics to build applicable formula or tools for measuring text readability for Vietnamese, a low- resourced language, to meet the demand for users and Vietnamese community in this era of technology.

References

Bùi, M. H. (2008). *Ngôn ngữ học Đối chiếu*. Hồ Chí Minh City, HCMC: Education Publishing House.

Cao, X. H., & Hoàng, D. (2005). *Từ Điển Thuật ngữ Ngôn ngữ học Đối chiếu Anh - Việt; Việt – Anh*. Hồ Chí Minh City, HCMC: Social Sciences Publishing House.

Cieri, C., Maxwell, M., Strassel, S., & Tracey, J. (n.d.). *Selection Criteria for Low Resource Language*. University of Maryland College Park, MD 20742, USA. Retrieved from https://pdfs.semanticscholar.org/315a/3a4a6db25e705f50159807917ec6f439f83b.pdf

Council of Euroupe (2010). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge Press. Retrievd from http://www.coe.int/en/t/dg4/linguistic/source/framework_en.pdf

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English, 26*, 23.

Dubay, H. W. (2004). *The Principles of Readability.* Impact Information, Costa Mesa, California.

Đinh, Đ. (2006). *Xử lý Ngôn ngữ tự nhiên*. Hồ Chí Minh City, HCMC: HCMC National University Publishing House.

Flesch, R. (1943). *Marks of a readable style, Columbia University contributions to education*, no. 897. New York: Bureau of Publications, Teachers College, Columbia University.

Flesch, R. F. (1948). A New Readability Yardstick. *Journal of Applied Psychology, 32*(3).

Flesch, R. F. (1949). *The Art of Readable Writing.* New York: Harper.

Flesch, R. F. (1974). *The Art of Readable Writing,* 25th anniversary edition, revised and enlarged. NY: Harper & Row.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh- Metrix: Providing Multilevel Analyses of Text Characteristics, *Educational Researcher*, *40*(5), 223-234.

Graesser, A.C., McNamara, D. S., & Louwerse, M. M. (2004). Coh- Metrix: Analysis of Text on Cohesion and Language, *Behavior Research Methods, Instruments, & Computers*, *36*(2). 193-202.

Gray, W. S., & Leary, B. E. (1935). *What Makes a Book Readable*. Chicago, Illinois: The University of Chicago Press.

Jamie, D. (2014). *Investigating the relationship between empirical task difficulty, textual features and CEFR levels*. EALTA 2014, 29 May – 1 June. University of Warwick

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. *CNTECHTRA Research Branch Report 8-75*.

Klare, G. R. (1973). *The Measurement of Readability*. Ames, Iowa: Iowa State University Press.

Lijun, F., Martin, J., Matt, H., & Noémie, E. (2010). *A comparision of Features for Automatic Readabiity Assesment,* Beijing August 2010, Poster Volume, 276-284.

Lorge, I. (1939). Predicting Reading Difficulty of Selections for Children. *The Elementary English Review*, *16*(6), 229-233.

Mai, N. C., Nguyễn Thị, N. H., Đỗ, V. N., & Bùi, M. T. (2007). *Nhập môn Ngôn ngữ học*. Hồ Chí Minh City, HCMC: Education Publishing House.

McLaughlin, H. (1969). SMOG Grading - a New Readability Formula. *Journal of Reading, 12*(8), 639-646. DOI: http://dx.doi.org/10.2307/40011226

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*, CUP.

Mostafa, Z., & Pooneh, H. (2012). Readability of Texts: State of the Art. *Theory and Practice in Language Studies*, *2*(1), 43-53. http://doi.org/10.4304/tpls.2.1.43-53.

Nguyễn, T. G., Đoàn, T. T., & Nguyễn, M. T. (2008). *Dẫn luận Ngôn ngữ học*. Hồ Chí Minh City, HCMC: Education Publishing House.

Viet Nam Committee of Social Sciences. (1993). *Ngữ pháp tiếng Việt*. Hanoi: Social Science Publishing House.

Vietnam Committee of Social Science (1993). *Vietnamese Grammar*. Hanoi: Social Science Publishing House.

Vu Thi, P. A. (n.d.). *Text Readability and testing languages*. Retrieved from http://ncgdvn.blogspot.com/2011/10/o-kho-cua-van-ban-va-viec-kiem-tra-ngon.html

Vu Thi, P. A. (2006). Khung trình độ chung Châu Âu và việc nâng cao hiệu quả đào tạo tiếng Anh tại ĐHQG – HCM. *Journal of Science and Technology Development*, *9*(10), 31-47.

Appendix

The parts of speech in Vietnamese calculated by CLC - Vietnamese – Tool kit, Computing Linguistics- CLC- University of Science Ho Chi Minh City

Label	Từ loại tiếng Việt (Vietmamese POS)	Ví dụ (Example)	Từ loại tiếng Anh (English equivalents)
Aa	tính từ hàm chất	lộ thiên, đầy, mắc	qualitative adjective
An	tính từ hàm lượng	đầu tiên	quantitative adjective
Cm	giới từ	giữa, của, trong, tại	major/minor conjunction
Cp	kết từ đẳng lập	cùng, với, và	parallel conjunction
Cs	kết từ chính phụ	nếu, thì, vừa, là	subordinating conjunction
D	phó động từ chỉ hướng	ra, vô, lên, xuống	direction
E	cảm từ	thưa, làm gì	emotion word
FW	từ nước ngoài	Miss, pH, super	foreign words
ID	thành ngữ	công ăn việc làm	idiom
M	trợ từ	đến, riêng, được, có, đó	modifier
Nc	danh từ đơn thể	bộ, ngôi, bản, con, bài	countable noun
Nn	danh từ	nước, người, chuyện, ông	common noun
Nq	danh từ số lượng	một vài, phần lớn, mấy	numeral
Nr	danh từ riêng	Tuấn, Hồng, Thành, Hà Nội	proper noun
Nt	danh từ chỉ thời gian	sáng, tối, năm, khi	temporal noun
Nu	danh từ chỉ đơn vị	TP., tỉnh, khu phố	concrete noun
ON	từ tượng thanh	tách, bùm bụp, hì hì	onomatopoeia
Pd	đại từ không gian, thời gian	nào, này, đó, bao giờ	demonstrative pronoun
Pp	đại từ xưng hô	tui, con, anh, chị, ông	personal pronoun
PU	dấu câu	Dấu phẩy, dấu chấm	punctuation
R	trạng từ	được, đều, chưa, nào	adverb
Vc	động từ so sánh	Là	comparative verb
Vd	động tử chỉ hướng	đến, ra, xuống	directional verb
Ve	động từ tồn tại	có, hết	state verb
Vv	động từ ý chí	viết, muốn, được, thay, ăn	volatile verb
X	không xác định	v.v	unidentified POS