Complex Word Identification in Vietnamese: Towards Vietnamese Text Simplification

Phuong Nguyen
Computer Science Department
Pomona College
phuong.nguyen@pomona.edu

David Kauchak
Computer Science Department
Pomona College
david.kauchak@pomona.edu

Abstract

Text Simplification has been an extensively researched problem in English, but has not been investigated in Vietnamese. We focus on the Vietnamese-specific Complex Word Identification task, often the first step in Lexical Simplification (Shardlow, 2013). We examine three different Vietnamese datasets constructed for other natural language processing tasks and show that, like in other languages, frequency is a strong signal in determining whether a word is complex, with a mean accuracy of 86.87%. Across the datasets, we find that the 10% most frequent words in many corpus can be labeled as simple, and the rest as complex, though this is more variable for smaller corpora. We also examine how human annotators perform at this task. Given the subjective nature, there is a fair amount of variability in which words are seen as difficult, though majority results are more consistent.

1 Introduction

Text Simplification is a task that focuses on improving the readability and understandability of text while preserving the original content and meaning. Text Simplification applications have been shown to benefit a variety of target audiences, including readers with low-literacy levels (Mason, 1978), non-native speakers (Paetzold, 2016), language learners (Gardner et al., 2007; Crossley et al., 2007), deaf people (Marschark and Spencer, 2010), people with reading comprehension problems such as aphasia (Carroll et al., 1998) and dyslexia (Rello et al., 2013), and people with Autistic Spectrum Disorder (Evans et al., 2014). It is also a useful preprocessing step for other NLP tasks, including parsing (Chandrasekar et al., 1996), information extraction (Evans, 2011; Miwa et al., 2010), and question generation (Heilman and Smith, 2010).

Although significant progress has been made in text simplification in multiple languages, including English (Coster and Kauchak, 2011; Nisioi et al., 2017; Woodsend and Lapata, 2011), Spanish (Saggion et al., 2015; Bott et al., 2012), Portuguese (Aluísio et al., 2008), Japanese (Katsuta and Yamamoto, 2019; Maruyama and Yamamoto, 2017), Korean (Chung et al., 2013), and Italian (Barlacchi and Tonelli, 2013), the problem remains a relatively new area of research in Vietnamese, a language spoken by over 70 million people (Van Driem, 2001) in Vietnam, the South East Asia region, France, Australia, and the United States. Sentence splitting has been conducted for the Vietnamese – English machine translation task (Hung et al., 2012), which can be helpful as an initial step for Text Simplification, but no further work has been recorded.

Other tasks in Vietnamese have been explored, from core problems such as dependency parsing, word segmentation, and part-of-speech parsing to more recent ones such as sentiment analysis, automatic speech recognition, and question answering.[1] Text Summarization is the most closely related task to Text Simplification that has been attempted in Vietnamese.

Progress on the specific task of Complex Word Identification in Vietnamese has not been reported so far. Although the terms *complex words* and *simple words* have appeared in literature on the Word Segmentation task, such as in Nguyen et al. (2006b), Nguyen et al. (2006a), and Anh et al. (2015), they refer to the length of each word (whether they are monosyllabic or polysyllabic words such as compound and reduplicative words) rather than the understandability and readability of each word in the context of Text Simplification.

We implement two approaches to solve the Complex Word Identification task in Vietnamese: frequency-based and classification-based with Support Vector Machines. We conclude with an experiment involving human annotators to predict the

[1] https://github.com/undertheseanlp/NLP-Vietnamese-progress

Proceedings of the Workshop on Multilingual Information Access (MIA), pages 59 - 68
July 15, 2022 ©2022 Association for Computational Linguistics

suitability of our datasets for this task.

2 Characteristics of Vietnamese

The characteristics presented in this section are extracted from Hạo (2000) and Hữu et al. (1998).

2.1 Language Family

Vietnamese is classified to be in the VietMuong group of the Mon-Khmer branch in the Austro-Asiatic language family.

Due to past colonization periods, Vietnamese is also heavily influenced by Chinese, as exemplified by the significant number of Sino-Vietnamese words (words with Chinese origin or consists of morphemes of Chinese origin) in the vocabulary, French, as seen in the use of calque (or loan translation), and English.

2.2 Language Type

Vietnamese is an isolating and tonal language with the following characteristics:

- It uses a Latin alphabet in conjunction with diacritics and several other letters.
- There are six tones marked by accents: level ("ngang"), falling ("huyền"), broken ("ngã"), curve ("hỏi"), rising ("sắc"), and drop ("nặng"). The pronunciation of these tones differ across the Northern, Southern and Central regions of Vietnam (Alves, 1995).
- It is a monosyllabic language.
- It is neither inflected nor conjugated, i.e. all words in Vietnamese are immutable.
- All grammatical relations are established by word order and function words.

2.3 A Word Unit

Vietnamese has a unit denoted "tiếng" that can represent either (Nguyễn et al., 2006):

1. a syllable with regards to phonology
2. a morpheme with regards to morpho-syntax
3. a word with regards to sentence constituent creation

Based on current literature, this unit is commonly referred to as a syllable. Thus, the Vietnamese vocabulary includes monosyllabic words ("từ đơn", words with a single syllable) or compound words ("từ phức", words with more than one syllable). About 85% of Vietnamese words are compound words and more than 80% of syllables are standalone words (Phuong et al., 2008; Dinh et al., 2008). This means that unlike in English and other Occidental languages that also utilize Latin alphabets, white spaces are not reliable indicators of word boundaries in Vietnamese. For example, "học sinh" (student) is a compound word that includes two syllables separated by a white space.

3 Data

We conduct two experiments across three Vietnamese corpora of various sizes extracted from different domains. We obtain a simple word list, a stopword list, and use the two lists to extract three complex word lists from the three corpora for evaluation purposes. The simple and complex wordlists for the three corpora are available online.[2]

3.1 Word Lists

The following two word lists are used:

- **Simple Word List**: A list of 3,000 words obtained by Luong et al. (2018) to construct a Vietnamese text readability formula. The list was used to replace the list of 3,000 words that fourth grade students can understand used in the Dale-Chall formula for English readability (Dawkins et al., 1956) in the development of an equivalent readability formula in Vietnamese.
- **Stopword List**: A list of 1942 stop words.[3]

3.2 Corpora

The following three corpora are used to conduct experiments. They are named according to the purpose of their construction.

- **READABILITY** (Luong et al., 2020)

  This corpus, constructed for research in Vietnamese text readability, contains 1,825 documents of approximately 3 million words in the literature domain. The documents were sourced from college-level textbooks, stories and literature websites, and were preprocessed for the minimization of spelling errors and standardization of punctuation, encoding, and

[2] https://github.com/phuongnguyen00/cwi-in-vietnamese

[3] https://github.com/stopwords/vietnamese-stopwords

tone. The corpus was then divided by experts into four categories: Very Easy (intended for children or people with middle-school education), Easy (intended for middle-school children or people with middle-school education), Medium (intended for high-school students or people with high-school education), and Difficult (specialized text intended for people with college education). Based on the Vietnamese Dictionary by Hoang (2017), more difficult groups of texts are more likely to include Sino-Vietnamese words and other words borrowed from English and French.

For this work we only use the Difficult subcorpus.

- **CLUSTER** (Tran et al., 2020)

  This dataset was constructed for the task of abstractive multi-document summarization. The dataset includes 600 summaries of 300 clusters with 1,945 news articles on five topics: world news, domestic news, business, entertainment and sports extracted from various news outlets aggregated by Google News in Vietnamese. Every cluster contains 4 - 10 articles, and the average number of articles per cluster is 6. Each document contains the following information: the title, the text content, the news source, the date of publication, the author(s), the tag(s), and the headline summary. These pieces of information are labelled using English.

  For this work we only use the original documents.

- **CLASSIFICATION** (Hoang et al., 2007)

  This corpus was constructed to solve the Text Classification task (labeling documents with a predefined topic). The corpus was comprised of articles from four major online newspapers, including VnExpress, TuoiTre Online, Thanh Nien Online, and Nguoi Lao Dong online. The data preprocessing phase included the removal of HTML tags, normalization of spelling, and other heuristics. There are 27 predefined topics ranging from music, family, and eating and drinking, to international business, new computer products and fine arts.

  The authors constructed 2 corpora of 2 levels of topic specificity (the higher level one included more fine-grained topic categorization). Corpus level 2 is used in this project.

3.3 Data Preprocessing

Since whitespace cannot be used to identify words in Vietnamese, we use the VNCoreNLP toolkit (Vu et al., 2018) for the word segmentation process. The word segmentation tool in the toolkit relies on the use of the Single Classification Ripple Down Rules (SCRDR) tree and was reported to achieve the best F1 score out of notable segmenters including vnTokenizer, JVnSegmenter, and DongDu (Nguyen et al., 2017).

We extract three complex word lists from the three corpora by removing all of the simple words, stopwords, proper nouns (words whose syllables are all capitalized), invalid words (such as words that contain numbers, letters, hyperlinks, and English words that are used repeatedly). The syllables in each word are concatenated with "_" as white spaces are not reliable indicators of word boundaries in Vietnamese. The remaining words are then identified as *complex*.

Table 1 shows various statistics for the three corpora. The Readability corpus has the smallest number of documents, but the documents tend to be longer. The Cluster corpus is the smallest of the three corpus with just over half a millions words. The Classification corpus is the largest, both in the number of documents and the number of words. These sizes are paralleled in the number of unique words from each corpora, though the Cluster corpus is high given its size indicating a slightly more difficult corpus. All of the corpora are comprised of about 60% simple words, though, again the Cluster corpus is slightly smaller than this.

For the experiments, we rely on the simple word list, and the 3 complex word lists as extracted above. We concatenate the simple word list with each of the 3 complex word lists to create 3 three separate datasets. These word lists will be referred to by their corpus' name in the following sections.

4 Methods

For each dataset, we have the simple word list and the list of unique complex words. This creates three complex word identification tasks to identify whether a word is simple or complex. We examine two approaches for the this task: frequency threshold and feature-based using Support Vector Machines.

	REA	CLU	CLA
docs	321	1945	25,286
words	1.58M	563K	4.96M
simple words	1.01M (64%)	315K (56%)	2.95M (59%)
stopwords	666K (42%)	174K (31%)	1.77M (36%)
unique complex words	10,273*	7,548	27,764

Table 1: Preliminary quantitative information of the three corpora. [REA = READABILITY, CLU = CLUSTER, CLA = CLASSIFICATION]

* *involves manual processing to remove foreign words and invalid words*

4.1 Frequency Threshold

For the Complex Word Identification task in English, frequency is an overpowering signal in determining whether a word is complex (Paetzold and Specia, 2016). The frequency approach only uses the frequency of a word in a particular corpus to label it as *complex* or *simple*.

For each of the three datasets that include both simple and complex words, we split it into training (75%) and testing (25%) data. Within the training dataset, we sort all of the words by frequency, and consider each frequency f out of all frequencies recorded as a cutoff point. For each frequency f, a word will be labelled complex if its frequency is smaller than or equal to f, and it will be labelled simple otherwise. We consider all possible frequencies f as the cutoff point and and identify the frequency that has the highest classification accuracy as our threshold for applying to the testing data.

4.2 Support Vector Machines Classifier

The frequency approach only utilizes a single feature. Many features have been suggested for use in the complex word identification task (Paetzold and Specia, 2016). For our classifier we used four features: corpus-specific frequency, number of syllables, number of characters, and number of characters and diacritics. All of the features besides word length try and capture different notions of word length. Some of these have worked well in other languages and some of these are specifically available in Vietnamese (i.e., diacritics).

The number of syllables is calculated based on the number of underscores found in a word. Because white spaces are not reliable indicators of word boundaries in Vietnamese, we concatenate the syllables of one word together with underscores in the data preprocessing step.

The number of characters and diacritics are calculated as the length of the word after being normalized into NFD (Normal Form D, also known as canonical decomposition)[4] with the `unicodedata` Python module.[5]

We used the `scikit-learn` package (Pedregosa et al., 2011) with the default regularization parameter $C = 1$ and the radial basis function kernel.

5 Experiments

We evaluate the performance of the two approaches on the three corpora based on overall accuracy and precision, recall, and F1 (for identifying simple words).

5.1 Frequency Threshold

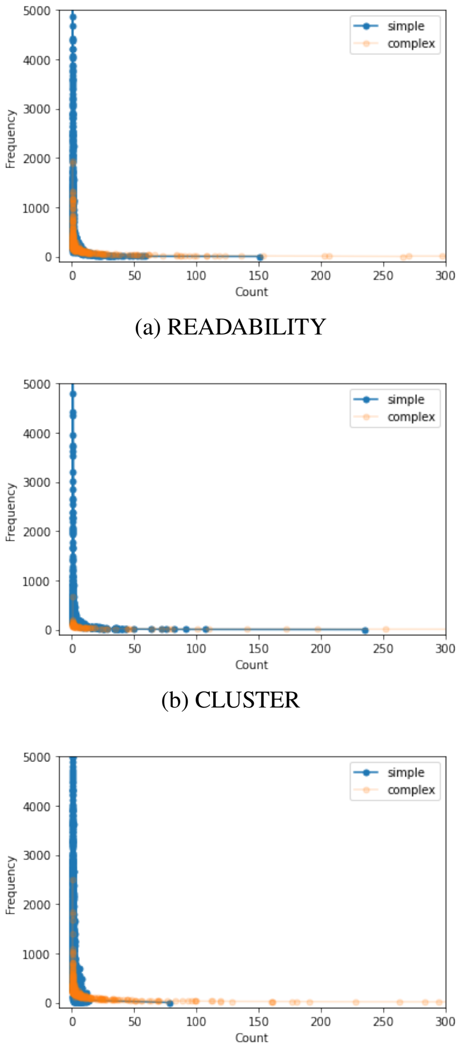
Frequency has been shown to be a strong signal in the CWI process. Figure 1 shows the frequency distribution of the three datasets. As expected, all three follow the standard Zipf's like distribution with a small number of words occurring very frequently and most of the words only occuring a small number of times.

Table 2 shows the accuracy, precision, recall and F1 scores. Overall, the approach does quite well with accuracies above 80% on all three corpora. The recall is high, highlighting that the approach is particularly good at identifying simple words. The results are significantly higher across all metrics on the Classification corpus. This is the corpus with the most data, and all documents represent news articles, which may have helped with consistency both because of source as well as writing practices.

Table 3 shows the cutoff frequencies and cutoff percentiles (if the words have frequencies below the percentile, then they are complex words). While cutoff itself varies significantly (mostly due to the size of the corpus), the percentage this frequency represents is much more consistent. For the two larger corpora, Readability and Classification, there is only a one percentage point difference: the top 10% most frequent words are the simple words. The Cluster dataset has a lower frequency cutoff.

[4] This method does not account for the diacritic found in the letter "đ", but accounts for all other diacritics.

[5] https://docs.python.org/3/library/unicodedata.html



(a) READABILITY

(b) CLUSTER

(c) CLASSIFICATION

Figure 1: The frequency distribution of the three full (unsplit) datasets.

We hypothesize this may have to do with its small size, though the source of the corpus might also play a role. More investigation is needed.

Figure 2 shows the accuracy distributions across possible cutoff frequencies for the three datasets. The pattern is consistent across the three datasets. The classification accuracy reaches a peak very quickly and then tends to taper off. The accuracy slightly drops and hits a plateau, except in the case of the Classification dataset in which the accuracy remains very high beyond the peak accuracy point.

5.2 Support Vector Machines Classifier

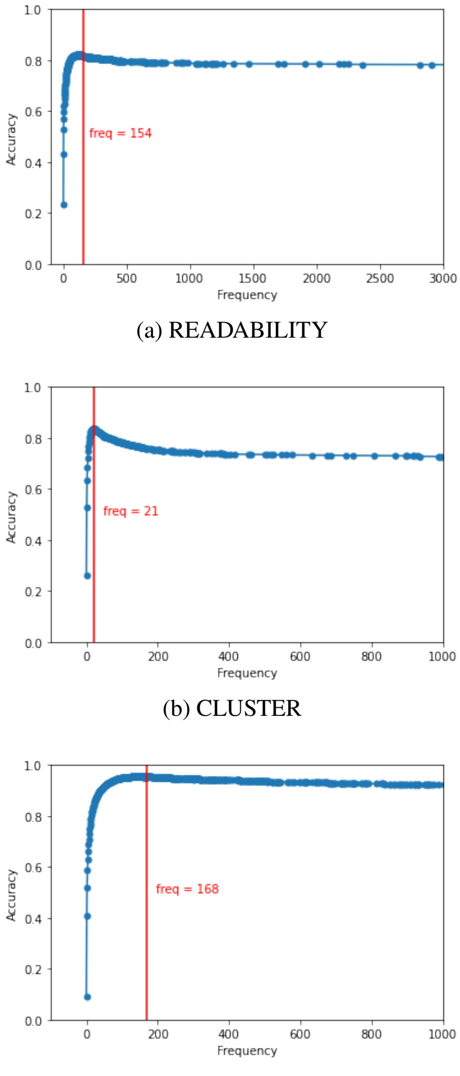
Table 4 shows the accuracy, precision, recall and F1 number for the feature-based SVM approach. The SVM approach tends to have slightly higher recall than the threshold approach, but the other met-

	accuracy	**precision**	**recall**	**F1**
REA	0.817	0.924	0.972	0.947
CLU	0.836	0.810	0.937	0.869
CLA	0.953	0.935	0.986	0.960

Table 2: The accuracy, precision, recall, and F1 scores of the Frequency Threshold approach across the three testing datasets. [REA = READABILITY, CLU = CLUSTER, CLA = CLASSIFICATION]

	cutoff frequency	**cutoff percentile**
REA	154	91.5%
CLU	21	79.6%
CLA	168	92.6%

Table 3: The cutoff frequency and the cutoff percentile of the three testing datasets. [REA = READABILITY, CLU = CLUSTER, CLA = CLASSIFICATION]



(a) READABILITY

(b) CLUSTER

(c) CLASSIFICATION

Figure 2: The accuracy distributions across possible cutoff frequencies of the three testing datasets.

	accuracy	precision	recall	F1
REA	0.821	0.820	0.983	0.894
CLU	0.825	0.821	0.967	0.888
CLA	0.954	0.958	0.992	0.975

Table 4: The accuracy, precision, recall, and F1 scores of the SVM classifier of the three testing datasets. [REA = READABILITY, CLU = CLUSTER, CLA = CLASSIFICATION]

	accuracy	precision	recall	F1
All	0.437	0.727	0.459	0.563
M	0.824	1.0	0.739	0.850

Table 5: The accuracy, precision, recall, and F1 scores of the human annotation process. [M = Majority]

rics are not significantly different. The additional features may provide some small information, but the SVM is still heavily relying on the frequency feature to make its prediction.

6 Human Annotation

To quantify the quality of the datasets for the automated CWI task in Vietnamese, three participants were asked to manually classify 199 words as simple or complex, with 100 words randomly picked from the simple words list and 99 words from the Readability complex word list. The words were presented by themselves without any additional context. All participants were native Vietnamese speakers pursuing a college degree in the United States. The instructions were provided in Vietnamese, in which an example of one simple word and one complex word is demonstrated. The participants were reassured that there are no right or wrong answers, encouraged to use their intuition when making the decision, and to label a word as complex when in doubt. Results are reported under two circumstances: a word gets assigned a label during this collective classification process if (a) the label is chosen by all 3 of the participants and (b) the label is chosen by a majority (i.e., 2 out of 3) participants.

Table 5 shows the results for the humans annotators. There is a drastic increase across all of the metrics when we remove the restriction that all annotators need to agree on a label. Accuracy increases two-fold from around 43% to 82%, and precision rises to 100%, meaning no simple words are mislabelled. Recall nearly reaches 75%, which reflects a decent level of agreement between the annotators' idea of complexity and what is represented in the Readability dataset. However, both between annotators as well as between the task construction, there is still some contention about which words are simple and complex. This highlights the difficulty and the subjectivity of this task.

7 Discussion

Frequency is an overpowering signal in determining whether a word is complex or simple as shown by the accuracy, precision, recall and F1 scores of the **Frequency Threshold** experiment, which are all are greater than 0.8 (see Table 2). Recall scores are all greater than 0.9 across the three datasets, indicating that this approach can reliably identify complex words. This finding is consistent with the results obtained from the Complex Word Classification task in English (Paetzold and Specia, 2016).

We analyze three corpora to try understand how consistent frequency is. For the larger corpora, it is surprisingly consistent with words in the top 10% most frequent words as simple. For smaller corpora this is more varied.

There are some shortcomings in the datasets that may affect the performance. There exist words in the simple word list that are acronyms that may be obvious to a certain target audience but not for the majority of Vietnamese readers (such as "UBND", which stands for "Uỷ ban nhân dân" (people's committee)), and can mean different things in different contexts (such as TP, which can mean "thành phố" (city) or "thành phần" (ingredient)). The Cluster and Classification datasets also involve foreign words, especially English words, that can add noise to the data.

Support Vector Machines are also explored to incorporate additional information into the prediction task. Three more features are added in addition to frequency for the SVM model: number of syllables, number of characters, and number of characters and diacritics. We hypothesize that longer words and words with more diacritics will be harder to recognize and understand. For example, "cỏ cây" (trees and plants) can be perceived as a simpler word to understand than "đường sá" (streets). However, results show that using SVM with more features do not improve the performance of the classification task compared to using a frequency threshold. In fact, we observe a decline in precision (from 92.40% to 81.95%) and F1 score (from 94.73% to 89.39%) on the Readability dataset. This can be explained by the fact that surface-level word features do not necessarily make the word more complex in terms of readability and understandability. Coming back to our example, although the former word "cỏ cây" is shorter and has fewer diacritics, it can also be simpler because both words have clear meanings ("cỏ" - grass and "cây" - plant), while the second syllable of the latter word "đường sá" is a Sino-Vietnamese word that may not be clearly decipherable. Because of this reason, "trung kiên" (loyal), which is a Sino-Vietnamese word, can be viewed as more complex than "phương hướng" (direction), which is a more common word. Again, this particular example shows that frequency gives a very strong signal.

The **Human Annotation** experiment shows a great difference between labeling based on the agreement between all three annotators or between the majority of annotators (2 out of 3 annotators). The accuracy and recall scores nearly double, and the precision score is 1.0 for the majority vote. This means that the majority of annotators' labeling of complex words is consistent with the data we obtain, which can indicate the suitability of the Readability dataset for the CWI training purposes.

8 Conclusions and Future Work

Several next steps can be taken beyond this project:

More Salient Features: Features that describe a word's characteristics beyond its pronunciation can be helpful to obtain a better classification performance. Some examples include sense count (number of entries in a dictionary for example), synonym count, and word type (whether the word is loan word).

Context: The approach we explore predicts words as simple/complex regardless of their context. In some cases, the context information can help provide additional information and additional features to help the identification (Paetzold and Specia, 2016).

More Diverse Human Annotators: Developing a clear definition of "word simplicity" and "word complexity" that reflects the needs of specific audiences by creating a bigger and more diverse pool of annotators with regards to gender, education background, and income level can also be helpful in constructing models that personalize text simplification for readers from different groups.

Text Simplification is the process of reducing the syntactical and lexical complexity of original text to make it more readable and understandable. Although this task has been shown to benefit various groups of audience and has been researched and experimented with extensively in English and several other languages, there has not been considerable progress made in Vietnamese-specific Text Simplification. In this study, we focus on the Complex Word Identification step in the Lexical Simplification pipeline, one approach to solve the Text Simplification problem. We view the question as a binary classification task, and conduct three experiments Frequency Threshold, Support Vector Machines, and Human Annotation to identify important features in the classification process and investigate the quality of our datasets for this particular purpose.

We observe that frequency is a very strong signal in the Complex Word Identification process in Vietnamese, shown by the Frequency Threshold experiment where we achieve a mean accuracy of 86.87% across our three datasets. The consistency of results across the three datasets give us a general rule to identify complex words in any corpus: the 10-20% of most frequent words are likely to be simple words. The use of Support Vector Machines with surface-level word features such as number of syllables and number of characters only marginally improves the recall scores but makes no significant difference in terms of accuracy, precision, and F1 scores. The Human Annotation experiment demonstrates how with a small number of annotators and a small sample, we can quantify how one dataset aligns with the definition of word complexity of college-educated native Vietnamese speakers. Considering the absence of significant progress on the Vietnamese-specific Text Simplification task and specifically the Complex Word Identification question, these three experiments constitute a first step in the exploration of the Lexical Simplification pipeline for Vietnamese.

References

Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248.

Mark Alves. 1995. Tonal features and the development of vietnamese tones. *Working Papers in Linguistics*, 27:1–13.

Tran Ngoc Anh, Nguyen Phuong Thai, Dao Thanh Tinh, and Nguyen Hong Quan. 2015. Identifying reduplicative words for vietnamese word segmentation. In *The 2015 IEEE RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF)*, pages 77–82. IEEE.

Gianni Barlacchi and Sara Tonelli. 2013. Ernesta: A sentence simplification tool for children's stories in italian. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 476–487. Springer.

Stefan Bott, Horacio Saggion, and Simon Mille. 2012. Text simplification tools for spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1665–1671.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer.

Raman Chandrasekar, Christine Doran, and Srinivas Bangalore. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pages 1–10.

William Coster and David Kauchak. 2011. Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation*, pages 1–9.

Scott A Crossley, Max M Louwerse, Philip M McCarthy, and Danielle S McNamara. 2007. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30.

John Dawkins, Edgar Dale, and Jeanne S Chall. 1956. A reconsideration of the dale-chall formula [with reply]. *Elementary English*, 33(8):515–522.

Quang Thang Dinh, Hong Phuong Le, Thi Minh Huyen Nguyen, Cam Tu Nguyen, Mathias Rossignol, and Xuan Luong Vu. 2008. Word segmentation of vietnamese texts: a comparison of approaches. In *6th international conference on Language Resources and Evaluation-LREC 2008*.

Richard Evans, Constantin Orasan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. Association for Computational Linguistics.

Richard J Evans. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and linguistic computing*, 26(4):371–388.

Elizabeth C Dee Gardner et al. 2007. Effects of lexical simplification during unaided reading of english informational texts. *TESL Reporter*, 40:33–33.

Eva Hasler, Adrià de Gispert, Felix Stahlberg, Aurelien Waite, and Bill Byrne. 2017. Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 45:221–235.

Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617.

Phe Hoang. 2017. *Từ điển Tiếng Việt (Vietnamese Dictionary)*. Da Nang Publising House.

Vu Cong Duy Hoang, Dien Dinh, Nguyen Le Nguyen, and Hung Quoc Ngo. 2007. A comparative study on vietnamese text classification methods. In *2007 IEEE international conference on research, innovation and vision for the future*, pages 267–273. IEEE.

Bui Thanh Hung, Nguyen Le Minh, and Akira Shimazu. 2012. Sentence splitting for vietnamese-english machine translation. In *2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 156–160. IEEE.

Cao Xuân Hạo. 2000. Tiếng việt-mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa (vietnamese-some questions on phonetics, syntax and semantics). *NXB Giáo dục, Hanoi*.

Đạt Hữu, TD Trần, and TL Đào. 1998. Cơ sở tiếng việt (basis of vietnamese).

Akihiro Katsuta and Kazuhide Yamamoto. 2019. Improving text simplification by corpus expansion with unsupervised learning. In *2019 International Conference on Asian Language Processing (IALP)*, pages 216–221. IEEE.

An-Vinh Luong, Diep Nguyen, and Dien Dinh. 2018. A new formula for vietnamese text readability assessment. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 198–202. IEEE.

An-Vinh Luong, Diep Nguyen, and Dien Dinh. 2020. Building a corpus for vietnamese text readability assessment in the literature domain. *Universal Journal of Educational Research*, 8(10):4996–5004.

Marc Marschark and Patricia Elizabeth Spencer. 2010. *The Oxford handbook of deaf studies, language, and education, vol. 2*. Oxford University Press.

Takumi Maruyama and Kazuhide Yamamoto. 2017. Sentence simplification with core vocabulary. In *2017 International Conference on Asian Language Processing (IALP)*, pages 363–366. IEEE.

Jana M Mason. 1978. Facilitating reading comprehension through text structure manipulation. *Center for the Study of Reading Technical Report; no. 092.*

Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. 2010. Entity-focused sentence simplification for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 788–796.

Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Minh Le Nguyen, and Quang Thuy Ha. 2006a. Vietnamese word segmentation with crfs and svms: An investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 215–222.

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2017. A fast and accurate vietnamese word segmenter. *arXiv preprint arXiv:1709.06307.*

Thanh V Nguyen, Hoang K Tran, Thanh TT Nguyen, and Hung Nguyen. 2006b. Word segmentation for vietnamese text categorization: an online corpus approach. *RIVF06.*

Thị Minh Huyền Nguyễn, Laurent Romary, Mathias Rossignol, and Xuân Lương Vũ. 2006. A lexicon for vietnamese language processing. *Language Resources and Evaluation*, 40(3):291–309.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, University of Sheffield.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Hông Phuong, Nguyên Thi Minh Huyên, Azim Roussanaly, Hô Tuòng Vinh, et al. 2008. A hybrid approach to word segmentation of vietnamese texts. In *International conference on language and automata theory and applications*, pages 240–249. Springer.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Advaith Siddharthan. 2011. Text simplification using typed dependencies: A comparision of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11.

Advaith Siddharthan and Angrosh Mandya. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 722–731.

Nhi-Thao Tran, Minh-Quoc Nghiem, Nhung TH Nguyen, Ngan Luu-Thuy Nguyen, Nam Van Chi, and Dien Dinh. 2020. Vims: a high-quality vietnamese dataset for abstractive multi-document summarization. *Language Resources and Evaluation*, 54(4):893–920.

George Van Driem. 2001. *Languages of the Himalayas: an ethnolinguistic handbook of the greater Himalayan region*, volume 2. Brill.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. Vncorenlp: A vietnamese natural language processing toolkit. *arXiv preprint arXiv:1801.01331.*

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification.

Transactions of the Association for Computational Linguistics, 4:401–415.