Pomona College
Department of Computer Science

Complex Word Identification in Vietnamese

Phuong Nguyen

April 21, 2022

Submitted as part of the senior exercise for the degree of
Bachelor of Arts in Computer Science

Dr. David Kauchak

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, the first step in the Lexical Simplification approach as defined by Shardlow [Sha13]. Our experiments across three datasets constructed for other Natural Language Processing tasks in Vietnamese show that frequency is a strong signal in determining whether a word is complex, with a mean accuracy of of 86.87%. From the consistency across the datasets, we deduce that 10-20% of most frequent words in any corpus can be labelled as simple, and the rest as complex. This project constitutes a first step in the exploration and implementation of the Lexical Simplification approach to simplify Vietnamese text.

Acknowledgments

I would like to thank Dr. David Kauchak for his support and guidance throughout this project. I would also like to thank Dr. Ami Radunskaya for her guidance and trust in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Text simplification (TS) is the process of reducing the linguistic complexity of text to increase its understandability and readability, while preserving its original meaning and content. The process may involve either lexical transformations, in which the vocabulary of the text is modified, or syntactic transformations, in which the structure of the sentences are modified [Sag18]. TS applications have been shown to be beneficial to people with limited literacy levels, non-native speakers, and people with various types of reading comprehension problems [ATA21].

Automated text simplification is a challenging problem that has been explored with different approaches since the late 90's. The progress made in this field follows the rapid development in statistical, machine learning, natural language processing and software techniques [ATA21]. TS remains an active research area with multiple unanswered or not satisfactorily addressed questions, such as personalized simplification, in which modifications for a specific target audience are performed, sentence joining, in which multiple sentences are concurrently considered, and improvement of evaluation metrics [AMSS20].

Vietnamese, the official language of Vietnam, is an Asian tonal language that is spoken by approximately 70 million people [VD01] worldwide, the majority of which are located in Vietnam. It is also the fifth most popular language spoken in the United States in 2010 with 1.38 million speakers, according to the US Census Bureau [Bur21].

Although significant progress has been made in Text Simplification in multiple languages, including English [CK11, NŠPD17, WL11a], Spanish [SŠB$^+$15, BSM12], Japanese [KY19, MY17], Korean [CMKP13], and Italian [BT13], Text Simplification remains a relatively new area of research with regards to Vietnamese. Sentence splitting has been conducted for the Vietnamese - English machine translation task [HLMS12], which can be helpful as an initial step for Text Simplification, but no further work has been recorded.

The benefits of Text Simplification for a variety of target audiences and the popularity of Vietnamese inspire the implementation this project. Because readers need to understand 95%[Lau89] − 98% [HCN00] of the vocabulary of a written text to be able to understand it, and 98% is a closer estimate for academic texts [SJG11], we choose to experiment with the Lexical Simplification approach (Refer to Section 2.1.1 for a more detailed description of this approach). We will implement two approaches to distinguish between simple and complex Vietnamese words: frequency-based and classification-based with Support Vector Machines. This is denoted the Complex Word Identification task, and it is normally seen as the first step in the Lexical Simplification pipeline. We conclude with an experiment involving human annotators to evaluate the quality of our datasets to solve the Complex Word Identification task.

1.1 Definitions

The terms *simple* and *complex* are frequently used to describe the nature of a text, and the appropriate label can vary depending on the context and the audience. Therefore, complexity and simplicity are relative concepts, and should be used with clear intention. The purpose of TS applications is to generate *simpler* (or *less complex*) than the original version [Sha14b].

Understandability and *readability* are two other significant terms to define as well, as they are used to explain the purpose of TS in its definition. While these two concepts are sometimes used interchangeably, they can refer to separate features of a text, depending on the context of an application. *Readability* describes how easy it is to read a text, and it is normally determined by the grammatical complexity, sentence length, and readers' familiarity with the vocabulary. *Understandability*, on the other hand, measures the amount of information that users obtain from reading a text, which is influenced by their familiarity with the vocabulary, their comprehension of the key concepts, and the effort put into reading the text. A text can have high *readability* but low *understandability* for a certain audience. For example, a scientific article can be well-written, but the density of domain knowledge may make it hard for readers without proper training to decipher. A text can also have low *readability* but sufficient *understandability* to be accessible to readers. An example is when an author conveys a simple idea with basic vocabulary by using confusing grammatical structure. *Readability* and *understandability* are thus closely related as the presence of both of them will make text more accessible.

1.2 Applications of Text Simplification

Simplified text could be beneficial for a variety of different types of readers, including people with low literacy levels, deaf people, people with autism, aphasia and dyslexia and non-native speakers. People with low-literacy levels are shown to find sentences with shorter structure easier to understand [Mas78]. Deaf children are shown to experience difficulty in understanding complex structures such as coordination, subordination and pronominalization [Q$^+$77], and passive voice and relative clauses [Rob81]. People with Autistic Spectrum Disorders (ASD) encounter difficulty in inferring contextual information and understanding long sentences with complex syntactic structures [EOD14]. People with aphasia see a decrease in the comprehensibility of a sentence following any increase in its grammatical complexity [She85]. People with dyslexia find reading more difficult if the words used are long and less frequent [RBYDMS13]. Simplified text is commonly used to teach beginner and intermediate English learners [CLMM07].

Motivated by the applicability of TS for a wide range of readers, research has been conducted to develop TS for specific audiences. For example, Paetzold and Specia [PS16b] used a context-aware word embeddings model and a corpus of subtitles to conduct lexical simplification for non-native English speakers. Orăsan et al. [OEM18] developed a TS software called OpenBook that can automatically identify a range of linguistic phenomena in a document that are potentially sources of confusion for people with high-functioning (IQ $>$ 70) ASD and replace some of them. Delvin and Unthank [DU06] built a web-based automated TS system that would make web content more accessible for people with aphasia by the simplification of vocabulary and syntax.

TS can also be a useful preprocessing step for other natural language processing tasks, including parsing [CDB96], information extraction [Eva11, MSMT10], question generation [HS10], summarization [SNM04], [SB12], [VSBN07], semantic role labeling [VK08], fact retrieval [KKM04], and machine translation [HdGS$^+$17]. TS has also been applied in medical research, such as for the simplification of medical literature [ODL$^+$07], drug package leaflets [SBM17], and patent documents [QKCH17].

While most of these applications are targeted at English speakers, the methods applied can shed light on the useful ways to tackle other languages as well.

1.3 Other Related Natural Language Processing Tasks

There are several related rewriting tasks that have different objectives and properties from those of TS.

Text summarization is a task that can be easily conflated with TS because both operations can reduce the complexity of the original text. However, Shardlow [Sha14b] defines text summarization to center around omitting unimportant or redundant information. Although deletion is allowed during TS, it's not the only operation. Other transformations allowed include replacement of terms with more explanatory phrases, addition of connectors to enhance flow of text, and explicit demonstration of co-connectors. Thus, the text's length can increase along with its readability and understandability.

Text compression focuses on reducing the length of text while retaining its main idea and grammaticality, and it can be integral to the text summarization process. Li et al. [LLWL13] introduced the concept of summary guided compression, which is a novel approach following the "sentence compression and sentence selection" pipleline of compressive summarization. This approach also tackles the **abstractive summarization** problem that focuses on summarizing the text as a whole with more sophisticated techniques. It's helpful to differentiate this approach with the **extractive summarization** technique of retention and concatenation of salient sentences in a text. The operations used in compressive summarization also include transformations such as substitution, reordering and insertion in addition to deletion, yet it is still a distinct process from text simplification because the principal goal is to shrink content rather than improving readability and understandability.

Split-and-rephrase [NGCS17] involves the splitting of a sentence into shorter ones and the necessary rephrasings to maintain meaning and grammaticality. As TS allows deletion, unimportant or peripheral information can be removed, which means the meaning of text is not completely preserved. Thus, split-and-rephrase can be perceived as one possible transformation technique within TS.

Chapter 2

Text Simplification

2.1 Simplification Approaches

There are four main categories to the TS problem: lexical, syntactic, monolingual machine translation, and hybrid techniques [ATA21]. The first three approaches are generally independent and methodologically different from each other. This section will define these four different techniques and illustrate them with specific studies.

2.1.1 Lexical Simplification

Lexical Simplification (LS) reduces the complexity of text through the identification and replacement of complex words with simpler ones. LS involves no modifications of the syntactic structure of a text and only focuses on simplifying the complex aspect of the vocabulary. The first Lexical Simplification system was proposed by Caroll et al. [CMC$^+$98], which simplifies English paper to support readers with aphasia. The system is comprised of an analyzer, which offers syntactic analysis, and a simplifier, which modifies the output of the analyzer to increase the readability of the text.

Shardlow [Sha14b] defines the following pipeline of four steps for LS:

1. **Complex Word Identification**: Detecting the complex words in a text that warrant simplification for a specific target audience

2. **Substitution Generation**: Producing a list of possible substitution candidates for the target complex word

3. **Substitution Selection**: Determining which element in the list of candidates that can replace the complex word and preserve both gram-

maticality and meaning of the sentence in its context

4. **Substitution Ranking**: Ordering of the selected candidates in terms of their simplicity in the given context

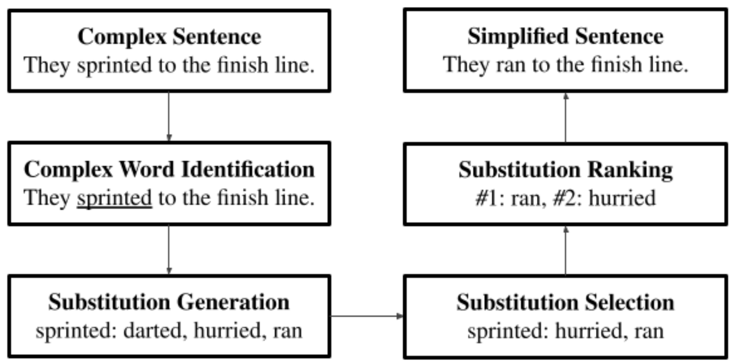


Figure 2.1: Lexical Simplification Pipeline

Figure 2.1 shows an example of the current pipeline, which follows what is presented in [PS17].

An extensive discussion of the first stage, Complex Word Identification, which is the focus of this project, is given in Section 2.2.

2.1.2 Syntactic Simplification

Syntactic Simplification (SS) involves the identification and transformation of complex grammatical structures in a text into simpler ones. Examples of syntactic phenomena that can be perceived as complex include subordination, coordination, relative clauses, or sentences that are not in the canonical word order [Sag18]. Syntactic simplification was introduced by Chandrasekar et al. [CDB96] who employed a rule-based method to modify sentences so that they could be correctly parsed by automatic systems. Their work set the foundation for current rule-based simplification approaches.

The majority of syntactic simplification approaches follow three stages [ATA21]:

1. **Structure and parse tree identification**: Words and phrases are clustered using "super-tags" that represent a part of the underlying sentence. "Super-tags" can be combined with conventional grammar rules

to provide a structured version of the text. During the analysis phase, the sentence's syntactical complexity is computed, which decides the necessity of simplification. This process can be automated with the use of matching rules or a binary classifier such as Support Vector Machine.

2. **Transformation**: Changes are made to the parse tree based on a set of rewrite rules. These rules specify the simplification operations, such as sentence splitting, clause reordering, and clause removal.

3. **Regeneration**: Further modifications are conducted to increase the cohesion, readability and understandability of the text.

SS studies follow one of the two directions: rule-based and data-driven. Most of current SS approaches are rule-based, and the performance of which is primarily dependent on linguistic expertise and accurate analyzing tools (parsers and taggers) [ATA21].

2.1.3 Machine Translation

TS has been viewed by several studies as a mono-lingual translation problem, where the source sentence in a complex language is translated into its equivalent in the corresponding simple language. This framing of TS as a machine translation (MT) problem is made possible by the availability of compara-ble and parallel corpora of original and simplified textual content. RecentMT-based attempts at TS utilize either the Statistical Machine Translation the Neural Machine Translation approach.

Statistical Machine Translation

Before the paradigm shift to Neural Machine Translation, Statistical Machine Translation (SMT) has been the mainstream approach to solve the translation problem. Given a sentence f in the source (complex) language, the objective of an SMT model is to generate a translation e in the target (simple) language. This is modeled by the noisy channel framework as follows:

$$e^* = \text{argmax}_{e \in E} p(e|f) = \text{argmax}_{e \in E} p(f|e)p(e),$$

where $p(f|e)$ is a translation model and $p(e)$ is a language model. A decoder is also used to produce the most probable translation e for an input f. In practice, the different implementations of the translation model and the decoder aim to maximize the translation quality rather than the generative noisy channel model.

Overall, SMT-based models' capabilities are limited to substitutions, short-distance reorderings and deletions. Without syntactic information or the addition of more expensive processes such as semantic analysis, they fail to produce quality splits [AMSS20].

Neural Machine Translation

Neural Machine Translation (NMT) is a recently proposed deep learning technique that relies on a single neural network to solve the translation problem [Sta20]. It has become the dominant paradigm in machine translation because of its more powerful capabilities compared to those of SMT systems.

2.1.4 Hybrid Approach

There exist several limitations to proposed TS solutions: for rule-based approaches, a substantial number of transformation rules is needed to achieve reasonable coverage, and for data-driven approaches, a parallel corpus is required. The generation and availability of these two resources may pose a significant challenge to the TS task. To overcome these issues, Siddharthan and Mandya [SM14] proposed a hybrid TS system that integrated a data-driven LS module with a hand-crafted rule-based SS module. The model was defined over synchronous dependency insertion grammars (SDIG), which provided an enhanced modeling of lexical transformations, simplified the rule generation step, and automated the acquisition of dependency parses from aligned sentences. The LS simplification module was trained on the EW-SEW alignment corpus. The SS module included the following transformation rules: 26 handcrafted rules for appositions and relative clauses, 85 rules for subordination and coordination, 11 rules for passive to active voice conversion, and 14 rules for standardization of quotations into the "X said Y" form. The results showed that this hybrid system surpassed a leading data-driven model at the time that used quasi-synchronous trees substitution grammar [WL11a] in terms of fluency, simplicity and meaning preservation.

2.2 Complex Word Identification

There are five categories in the first stage of the Lexical Simplification pipeline: Simplify everything, Threshold-based, Lexicon based, Implicit CWI, and Machine learning-assisted [PS17].

2.2.1 Simplify everything

Early LS approaches [Dev98] skipped the CWI step because all words in a sentence were assumed to be simplifiable. This method has proven to be not effective: Shardlow [Sha14b] and Paetzold [PS13] demonstrated that a simplifier without a CWI module might replace words which are already easy to comprehend by the targeted audience, and hence could complicate the text even more with out-of-context word choices or ungrammatical phrasings after the simplification process.

2.2.2 Threshold-based

The objective of threshold-based approaches is to find a threshold t over a metric of simplicity M for a word w such that if $M(w) < t$, then the word w can be labeled as either simple or complex.

Word length is one example of a simplicity metric used in the CWI step. Keskisärkkä [Kes12] described an LS approach in which complex words were replaced with their most frequent synonym. Results indicated that an increase in the word length decision threshold led to a decrease in the number of errors. In other words, simplifying words with more than 7 letters produced sentences with higher readability scores than simplifying all words in a sentence.

Word frequency has been a more popular choice for threshold-based LS approaches. Leroy et al. [LEK$^+$13] proposed an approach in which only words with frequency count of less than around 15 million times are simplified. This threshold was chosen because it was the occurrence count of the 5000th most frequent word in the Google 1T corpus. Based on human evaluation, the reading difficulty of the text produced by the system was significantly reduced.

Although threshold-based approaches are intuitive and relatively straightforward to implement, their applicability in practice is questionable. Shardlow [Sha14a] conducted with the goal of discovering the most frequent types of errors made by a baseline LS approach. The manual evaluation of the model used in the study, which relied on the Kucera-Francis coefficient [Rud93] for the identification of complex words, showed a 65% error rate in the identification step, the majority of which is the mislabelling of simple words as complex.

2.2.3 Lexicon-based

The lexicon-based approach was developed to address the limitations of the threshold-based one. This strategy identifies a simplifiable word using the lexicon of complex words: if a word w is a part of the lexicon of complex words L, then it is labeled as simplifiable [PS17].

The FACILITA system developed by Watanabe [WJU$^+$09] is an example in which the lexicon-based approach was applied successfully. The tool was designed to simplify web pages as part of the PorSimples project [AG10], a simplification framework for low literacy readers of Portugese. The lexicon used to detect simple words was derived from children books, a list of frequent words in news documents and a set of words chosen by human judges. FACILITA was shown to effectively support its target audience to consume texts of complex nature, such as news articles.

There are several limitations to the lexicon-based approach, including the high cost to create large lexicons of complex and simple words, and the absence of a universal complexity scale that applies to every possible target audience.

2.2.4 Implicit Complex Word Identification

More recent LS approaches incorporate the CWI step implicitly in the remaining steps of the pipeline rather than explicitly conduct CWI as an initial step. In these approaches, all words were viewed as complex, and during the simplification process, substitutions where a word is replaced with a more complex counterpart were removed.

Bott et al. [BRDS12] defined a word simplicity metric that is based on word frequencies and word length. This metric was used to exclude candidate substitutions which are evaluated to be more complex that the original word. Glavaš and Štajner [GŠ15] used a similar approach, in which a target word was replaced only if it had lower frequency than that of the selected substitution candidate.

Viewing CWI as an implicit step in the LS moves the focus on the categorization of words as inherently complex or simple to the finding of simpler substitutions. This method can be especially helpful in the cases where the training data is seen as capturing the needs of the target audience [PS17].

2.2.5 Machine learning-assisted

Machine learning techniques can be used to learn a model of word complexity. With a training set of words that are labeled as either complex or simple, the CWI stage can be viewed as a binary classification problem. If the labels are complexity quantifiers, regression techniques can be employed to quantify the level of complexity of a given word.

Shardlow [Sha13] compared the performance of a Support Vector Machine (SVM) classifier, a threshold-based strategy and the "simplify everything" approach.

The following features were identified for classification with SVM: (1) word frequency in the SUBTLEX corpus [BN09], which were comprised of over six million sentences extracted from movie subtitles, (2) CD count (number of films in which a word appeared in the SUBTLEX corpus), (3) length (number of characters), (4) syllable count, (5) sense count (the number of ways a word can be interpreted), and (6) synonym count. The last two features indicate the level of ambiguity of a word, and were extracted from WordNet. Other potential features not employed include lexical contextual information (simple words tend to be used along with other simple words) and the surrounding syntax (the complexity of the syntactical structure can correlate with that of vocabulary used). The RBF kernel was used, and the 2 parameters C and γ were selected by grid search. The SVM model was trained and tested with the CW corpus, which contained Wikipedia sentences with a single target complex word and a simpler alternative.

Results showed that SVM achieved the highest precision (the ratio of the number of correctly predicted words over the number of all words identified as complex, or the number of true positives over the sum of true positives and false positives) out of the three models. However, it attained the lowest recall (the ratio of complex words correctly identified over all complex words, or the ratio of true positives over the sum of true positives and false negatives - see Subsectioni 5.3). The "simplify everything" approach achieved the best score because of the assumed perfect balance between the number of complex and simple words present, which was not representative of a typical CWI task, in which a significant discrepancy between the number of complex and single words is usually observed.

2.2.6 Performance Comparison

Paetzold and Specia [PS17] showed that with labeled datasets, supervised approaches that use highly tuned modern machine learning techniques are likely to outperform other alternatives. On the contrary, Malmasi et al. [MDZ16] and Konkol [Kon16] demonstrated that although resource-heavy models normally attain the best performance scores in CWI, effective systems can still be developed with resource-light approaches and without heavy tuning. In situations with no training data available, lexicon-based approaches can be utilized. Although these approaches are context-specific, meaning they are normally not generalizable for different target audience, they can be useful in scenarios where certain assumptions about the source vocabulary can be reasonably made. For example, documents in the medical domain will include technical terms that are unfamiliar to the non-expert reader.

2.3 Challenges and Future Directions

Text Simplification has come a long way since being solved with the extractive approach, where only sentences with the “most" meaning are retained in a paragraph or document. Most of the current approaches in TS are abstractive, in which operations such as sentence splitting, and text deletion and addition are allowed [SM20]. The four main simplification approaches of Lexical Simplification, Syntactic Simplification, Machine Translation, and Hybrid have been experimented with extensively in different languages by different research groups around the world. Although these approaches all attain promising results and reveal important findings, there are several challenges in TS that are worth addressing in the future.

Textual Dimension. Most TS models focus on sentence simplification, and research on tackling the TS problem from the document-level perspective is scarce. By treating TS as a sentence simplification problem, critical aspects of cohesion, coherence and style are disregarded. For example, in Lexical Simplification approaches where a word is replaced with its synonym, in gendered languages such as Spanish, focus needs to be placed on not only local agreement issues, which involves the replacement of adjectives or determinants which could alter the word in question, but also long-distance agreement issues [Sag18]. Woodsend and Lapata [WL11b], and Mandya, Nomoto and Siddharthan [MNS14] produced sentence-level simplifications and conducted document-level readability optimization. However, Saddiharthan claimed that syntactic changes to sentences (especially splitting) could impact the rhetorical relations between them, which could only be rectified by looking beyond sentence boundaries. This remains an exciting area of research as document-level simplification best approximates the ideal use case of text simplification. This line of research may start with the identification of the differences between document simplification and sentence simplification as alterations that cover multiple sentences are likely to be necessary when the original text is viewed as a whole rather than as a collection of sentences. Furthermore, the curation of proper corpora for training and testing purposes and the devise of new evaluation metrics are necessary to advance this approach.

Corpora Breadth and Depth. More work needs to be done regarding the depth and breadth of the datasets used for TS: the diversity and accessibility of high-quality datasets within one language, and the availability and substantiality of datasets across different languages. Regarding depth for English as a high-resource language, the majority of datasets used in TS research is based on English Wikipedia and Simple English Wikipedia (EW-SEW) as these datasets are publicly available and provide automatically collected alignments between sentences of equivalent articles. There exist several variations of EW-SEW datasets based on the different methods of alignments extraction, all of which suffer from certain limitations, including the presence of noise (misalignments), the limited size, and the limited variety of alignments (for example, only having 1-1 alignments) [AMSS20]. The Newsela corpus [XCBN15] was created by professionals to resolve some of the problems introduced by EW and SEW datasets, and thus it is a higher-quality dataset. However, the fact that the common splits of the data cannot be shared publicly impedes the development and objective comparison of models that utilize it. TS would benefit from high-quality, professionally produced, publicly distributable datasets that combine the strengths of both EW-SEW and Newsela datasets. An increase in the diversity of datasets in terms of application domains, target audience and text transformations applied is also important in the advancement of TS. Lastly, during the evaluation process, the collection of various simplification references per simplified output, as done by Xu et al. [XNP$^+$16], is also a desirable practice.

Regarding breadth across different languages, the availability and accessibility of datasets in languages other than English could be helpful. For example, while there exists some parallel simplification datasets in Brazilian Portuguese, their sizes may not be sufficient to be used in machine learning approaches [Sag18]. Although general-purpose lexical resources such as WordNet have been employed in TS to obtain synonyms, these resources are not the most well-suited for the simplification task as they do not include readability information that may be necessary during the simplification process [FGWF14].

Target-Specific and Personalized Simplification. Most of current TS research is centered on the learning process of the simplification operations used in the training corpora rather than the needs of a specific target user (with a few exceptions, such as Saggion et al.'s [SFS+17] work on creating an accessible email client for people with intellectual or developmental disabilities, and Rello et al.'s [RBYDMS13] research on text simplification strategies for people with dyslexia). The negligence of a specific target audience during the development process can create a simplification system that either includes unnecessary transformations that undermine readers' capabilities, or miss important treatments of complex linguistic phenomena. Furthermore, most simplification approaches do not rely on a model of a specific target audience' lexicon that could be of use during the simplification process. The construction of modular, customizable systems that are adaptable to the needs of different types of users is an important goal for the research community in TS. Taking target-specific development a step further, as individuals within the same audience group may have specific simplification needs and preferences, a model that can learn from its interactions with users for a personalized experience would be useful.

Explanation Generation. Although the majority of TS approaches do not focus on specific simplification transformations, they mostly address four main operations: deletion, substitution, reordering and splitting [AMSS20]. Nevertheless, by definition, TS can also involve the addition of information that clarifies complex terms or concepts. This operation does not purely substitute a complex expression with a simpler one or its definition. Instead, it elaborates on a challenging concept that preserves the grammaticality and meaning of the sentence while improving its simplicity. Only limited work has been published on explanation generation for TS, including that of Eom et al. [EDS12], Kandula et al. [KCZT10] and Watanabe et al. [WJU+09].

Simplification Evaluation. Regarding automatic evaluation, there exist only two simplification specific metrics: SARI (System output Against References and Input sentence) [XNP+16], which focuses on paraphrasing, and SAMSA (Simplification Automatic evaluation Measure through Semantic Annotation) [SAR18], which focuses on sentence-splitting. However, humans perform several more transformations that are not reflected in the evaluation process of a model's output. Other content-based evaluation metrics such as BLEU [PRWZ02] borrowed from automatic translation have also been used to evaluate the performance of an output of a simplification system. Improving the automatic evaluation and comparison of different TS models are necessary. Research on Quality Estimation has shown promising results on the use of reference-less metrics for evaluation that can improve the speed and scale of automatic assessment. This branch of work has been applied in several studies, including Štajner et al.'s [ŠFSRP18] work on a language-independent sentence alignment system from comparable TS sources and Martin et al.'s [MHM+19] comparison of multiple approaches to reference-less quality estimation of sentence-level TS models.

Regarding human evaluation, there are three main criteria used to assess an output of a TS system: (i) meaning preservation (i.e. is the meaning of the simplified text equivalent to that of the original text?), (ii) grammaticality (i.e. are the simplified sentences correct?), and (iii) simplicity (i.e. is the output simpler than the original?). Are these criteria sufficient for evaluation? Would these criteria remain relevant for document-level simplification approaches? For specific-audience approaches? These questions are also important to tackle the TS problem.

Chapter 3

Vietnamese as a low-resource language

3.1 Characteristics of Vietnamese

As the official language of Vietnam, Vietnamese is the primary language used in media outlets and in the education system throughout the country. Vietnamese is also the 21st most spoken in the world [Anh21]. With over 70 million speakers [VD01], Vietnamese is spoken in Vietnam, the South East Asia region, France, Australia and the United States. The following characteristics are mostly referenced from Cao [H0], Doan [Doa99], and Hữu et al. [ĐDL98], unless otherwise noted.

3.1.1 Language Family

For a long time, linguists did not reach a consensus on the genetic affiliation of Vietnamese [Dif89]. Vietnamese has been affiliated with Chinese [Tab38] and Tai [Mas12] in the past. However, since the work of Haudricourt, scholars now classify Vietnamese in the VietMuong group of the Mon-Khmer branch in the Austro-Asiatic language family.

Vietnamese uses a Latin alphabet in conjunction with diacritics and several other letters. Due to past colonisation periods, the Vietnamese language was 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.

3.1.2 Language Type

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

- 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 [Alv95].
- 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.

3.1.3 Vocabulary

A Word Unit

Vietnamese has a unit denoted "tiếng" that can represent either:

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 stand-alone words [PTMHR+08, DLN+08]. For example, both syllables in "nhà cửa" (houses) can function independently. However, there exist compound words such as "bỡ ngỡ" (bewildered) whose syllables do not necessarily all carry a meaning.

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.

Compound Words

Within compound words ("từ phức"), there are several subcategories:

1. Compound words formed by phonetic reduplication ("từ láy")
   This phonetic reduplication serves one of the following functions:

   - It can enhance the meaning of the word. For example, "chật chội" indicates a narrower space compared to the core syllable "chật", and "sạch sành sanh" emphasizes on the absolute state of "sạch" (clean).
   - It can dampen the intensity of the meaning of the word. For example, "đo đỏ" depicts a lighter red than "đỏ".
   - It can imaginatively demonstrate the repetitiveness of an action. For example, "rung rung" is more imaginative than "rung" (shaky).
   - It can signal incontiguous but cyclic nature of something. "Lấp ló" (can be seen from a far but not clearly) or "lập loè" (flicker) are two examples.
   - It can mark a perfect state of things. Examples include "ngay ngắn" (organized) or "vuông vắn" (organized in a specific way).

   There are several notable facts of this group of compound words:

   - The syllables do not have to start with the exact characters, but there must be a consistency in terms of pronunciation. For example, "cuống quýt" (hurry) has different starting consonants "c" and "qu", but they are pronounced in a nearly identical way, so the word is categorized in this group.
   - It is not always clear which syllable is the primary one because all syllables do not necessarily have a meaning on their own. Examples of this include the words "nhí nhảnh" (playful, joyful) or "bâng khuâng" (undecided or melancholic).
   - Words with identically pronounced starting consonants do not necessarily belong to this group. For example, "đi đứng" (walk) belongs to another group of compud words.

2. Compound words formed by semantic coordination ("từ ghép đẳng lập")
   All syllables in these words contribute equally to the meaning of the

word. They can either stand independently or there exists one or more components that have lost its meanings . For example, "bố" (father) and "mẹ" (mother), two stand-alone words, can be combined into "bố mẹ" (parents). Another example is "xe cộ" that refers to general traffic. "Xe" means "vehicle", but "cộ" does not imply any meaning.

3. Compound words formed by semantic subjugation ("từ ghép chính phụ")
   In these words, there is one core syllable that typically precedes every other syllable. Normally, the core syllable indicates a category of objects, and the other syllables add a layer of specificity. For example, in the word "mùa xuân" (spring), "mùa" means season and "xuân" indicates which season it is. This phenomenon resembles hypernyms and hyponyms in English.

Loan Words

Vietnamese vocabulary is influenced by Chinese, English and French [Alv09]. Words loaned from these foreign languages can be classified as a special type of Compound Words.

Chinese is the chief source for Vietnamese loan words due to China's thousand-year domination of Vietnam . Some examples of Sino-Vietnamese words are "quốc gia" (nation), "định cư" (settle) and "bình minh" (sunrise). Certain Sino-Vietnamese words are so common that it may be challenging for Vietnamese speaker to identify as loan words.

Starting from the end of the 19^{th} century, French colonization showed an impact on Vietnamese lexicon. The loan words from French included terms related to clothing, food, household goods, and technological inventions that reflect the socio-cultural influence of France on Vietnam. For example, "bơ" (butter , or beurre in French), "cà phê" (coffee, or café in French), "măng tô" (cloak, or manteau in French), and "ga" (train station, or gare in French).

Since the 1960s, because of the American presence in Vietnam, loan words from English are added to Vietnamese vocabulary because of the lack of equivalent terms. For example, "ti vi" (TV) and "top" (the top position in a list) are borrowed from English. Because of the increasing global ubiquity of English, multiple English words are used directly in its original form without being converted to Vietnamese using the official alphabet, such as "wifi" (the Vietnamese alphabet does not have the letters "w" and "f") and "internet".

Idioms

Idioms ("thành ngữ") are immutable memorable short phrases that express a socio-cultural phenomenon that should be interpreted beyond the literal constituting components. A few examples include "thuận buồm xuôi gió" (smooth sailing), "câm như hến" (as silent as a mussel), and "cưng như cưng trứng, hứng như hứng hoa" (which means the meticulous handling of an object or gentle behavior towards a person). Although idioms normally consist of 3 or more syllables, because they do not form a complete sentence, idioms are normally viewed as one word. For our classification purpose, all idioms are viewed as compound words as well.

Idioms can be easily confused with proverbs ("tục ngữ"). However, proverbs signify a complete idea and is typically seen as a full sentence.

For a more comprehensive linguistic description and characterization of Vietnamese, interested readers may reference [Ngu97], [Tan07] and [Alv06].

3.2 Related Work in Vietnamese

Although Vietnamese is a low-resource language, significant progress has been made on multiple NLP tasks in the language, 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. State-of-the-art results and datasets of different tasks are recorded in a GitHub repository of Under The Sea, a Vietnamese NLP research group.

Text Simplification is not listed as a task on this repository. The most closely related task is Text Summarization, the difference between which and Text Simplification is explained in Section 1.3. Nguyen et al. [NNN$^+$18] implemented and compared multi-document summarization approaches in three categories: unsupervised, supervised and deep learning on two datasets, each of which contains articles from Vietnamese online news outlets divided into 200-300 topics. Results showed that Multi Additive Regression Trees (MART), one of the learning-to-rank methods based on gradient boosted regression trees, achieved promising results and even outperformed unsupervised learning methods when evaluated with ROUGE-scores on various lengths of references.

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 [NTNN06], [ATTQ15], and [NNLNH06], they refer to the length of each word (whether they are monosyllabic or polysyllabic words, i.e. compound and reduplicative words – see Section 3) rather than the understandability and readability of each word in the context of Text Simplification.

Chapter 4

Data

We conduct two experiments across three Vietnamese corpora of various sizes extracted from different domains. We obtain a Simple Word List, a Stop Word List, and use the two lists to extract three Complex Word Lists from the three corpora for training purposes.

4.1 Word Lists

The following two word lists are used:

- **Simple Word List**: A list of 3000 words obtained by Luong et al. [LND18] to construct a Vietnamese text readabillity formula. The list was used to replace the list of 3000 words that fourth grade students can understand used in the Dale-Chall formula for English readability in the development of an equivalent readability formula in Vietnamese.

- **Stop Word List**: A list of 1942 stop words. [1]

4.2 Corpora

Three corpora are used to create the datasets used in the experiments:

- **READABILITY** [LND20]

  This corpus, constructed by Luong et al. for research in Vietnamese text readability, contained 1825 documents of approximately 3 million words in the literature domain. These documents were sourced from

[1] This Stop Word list is publicly available on Github

college-level textbooks, stories and literature websites and were preprocessed for the minimization of spelling errors and standardization of punctuation, encoding, and 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 children or people with high-school education), and Difficult (specialized text intended for people with college education). Based on the Vietnamese Dictionary by Hoang [Hoa17], more difficult groups of texts are more likely to include Sino-Vietnamese words and other words borrowed from English and French.

In this project, only the Difficult sub-corpus is utilized.

- **CLUSTER** [TNN+20]

  This dataset is constructed by Tran et al. for the task of abstractive multi-document summarization. The dataset includes 600 summaries of 300 clusters with 1945 news articles on five topics: world news, domestic news, business, entertainment and sports extracted from various of news outlets aggregated by Google News in Vietnamese. Every cluster contains 4 - 10 articles, and the average number is 6 articles per cluster. 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.

  In this project, only the original documents are utilized.

- **CLASSIFICATION** [HDLNN07]

  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.

4.3 Data preprocessing

4.3.1 Word Segmentation Tool

In this project, the VNCoreNLP toolkit [VNN+18] is used for the word segmentation process. VNCoreNLP is an open-source Natural Language Processing pipeline for Vietnamese that can efficiently and reliably perform the key NLP tasks of word segmentation, part-of-speech tagging, named entity recognition, and dependency parsing. 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 F−1 score out of notable segmenters, including vnTokenizer, JVnSegmenter, and DongDu [NNV+17].

4.3.2 Data preprocessing

We extract three complex word lists from the three corpora following two steps: (1) word segmentation carried out by the corresponding tool in the VnCoreNLP toolkit and (2) removal of simple words, stop words, proper nouns, 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 complex word list extracted from the READABILITY corpus involves some further preprocessing because it includes words in multiple other languages, such as French, English, Chinese and Russian.

We first experiment with three available language detection packages: `langdetect`, a package ported from Google's language detection tool, `spacy-langdetect`, a fully customizable language detection pipeline, and `fasttext`, a tool supporting text-based language identification. However, these tools misclassify around 10% of words as non-Vietnamese. Closer inspection reveals that words without diacritics such as "mong manh" are likely to be misclassified and removed, which can send an inaccurate signal regarding the importance or the lack thereof of the presence of diacritics in the classification process.

Therefore, after manually removing foreign words and other invalid words (such as missegmented words by the segmentation tool and typos), the list is downsized from approximately 14K words to 10K words and better reflects the Vietnamese language.

Further quantitative information of the three corpora and their corresponding complex word lists are provided in Table 4.1.

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

	READABILITY	**CLUSTER**	**CLASSIFICATION**
Document Count	321	1945	25,286
Word Count	1,577,683	563,306	4,962,725
SW Count	1,007,392 (63.85%)	314,546 (55.84%)	2,951,129 (59.47%)
Stop Word Count	665,527 (42.18%)	174,427 (30.96%)	1,772,425 (35.71%)
Unique CW Count	10,273*	7,548	27,764

Table 4.1: Preliminary quantitative information of the three corpora READABILITY, CLUSTER, and CLASSIFICATION. [SW = Simple Word, CW = Complex Word]

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

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.

Chapter 5

Methods

Our attempt at the Complex Word Identification task involves two experiments: frequency threshold and binary classification with Support Vector Machines.

5.1 Frequency Threshold

We learn from the Complex Word Identification in English task that frequency is an overpowering signal in determining whether a word is complex [PS16a]. The frequency threshold experiment involves only using 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 then calculate the accuracy of using f as the cutoff. After trying all possible frequencies f as the cutoff point and calculate the corresponding accuracy, we record the f that has the highest classification accuracy as our threshold for the testing data. We then report the accuracy, precision, recall and f-1 scores of the classification process on the testing data (refer to Section 5.3 for more information about these metrics).

5.2 Support Vector Machines

We aim to use Support Vector Machines (SVM) to improve the results obtain by the Frequency Threshold approach. The features used are corpus-specific frequency, number of syllables, number of characters, and number of characters and diacritics. The features other than frequency are chosen because we hypothesize that longer words are more likely to be complex.

The number of syllables are 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) [1] with the `unicodedata` Python module [2].

We rely on the Support Vector Classifier implementation provided in the `scikit-learn package` [PVG+11] for our classification task [3]. We evaluate its performance with accuracy, precision, recall and f-1 (refer to Section 5.3 for more information about these metrics).

5.3 Metrics

We first define the following terms in the context of the binary classification problem of complex word identification:

- **True positive** (TP): A complex word correctly labelled as complex by the classifier
- **True negative** (TN): A simple word correctly labelled as simple by the classifier
- **False positive** (FP): A simple word incorrectly labelled as complex by the classifier
- **False negative** (FN): A complex word incorrectly labelled as simple by the classifier

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

[2] The information on the `unicodedata module` can be found on the documentation website for Python

[3] The implementation details of the SVC module is on the `scikit-learn website`

Four metrics are used to evaluate the performance of the approaches: accuracy, precision, recall, and f-1 score. They are defined as follows:

- **Accuracy** is an intuitive metric that shows the ratio of correctly identified words in both complex and simple classes over all words.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5.1}$$

- **Precision** shows the ratio of truly complex words out of all words labelled as complex. High precision means a low false positive rate.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.2}$$

- **Recall** shows the ratio of words identified as complex by the model out of all complex words in the dataset.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.3}$$

- **F-1** score provides a harmonic mean of precision and recall.

$$\text{F-1} = \frac{2 \cdot (\text{Recall} \cdot \text{Precision})}{(\text{Recall} + \text{Precision})} \tag{5.4}$$

Chapter 6

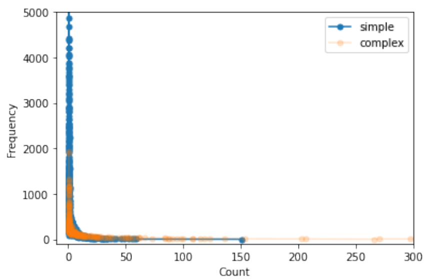
Experiments

We approach the Complex Word Identification task with two approaches: frequency threshold and Support Vector Machines classification using 4 word features: corpus frequency, number of syllables, number of characters, and number of syllables and diacritics. The use of SVM does not significantly improve the results obtained from the frequency-based approach.

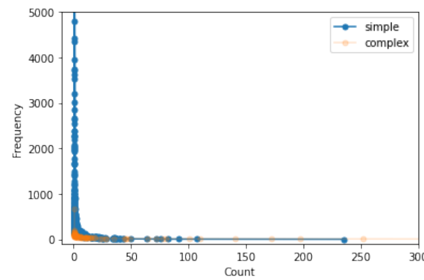
6.1 Frequency Threshold

Frequency has been shown to be a strong signal in the CWI process. With each of our word lists, we first obtain the frequency of each word from its corresponding corpus. We then split the dataset into testing (25%) and training data (75%). Figure 6.1 shows the frequency distribution of the three datasets. Finding the frequency threshold is similar to finding a horizontal line that best splits the two classes of complex and simple words. For each frequency f in our corpus-specific training data, a word w is classified as complex if its frequency is less than or equal to f, and simple otherwise. The frequency with the best accuracy is used on the testing data.

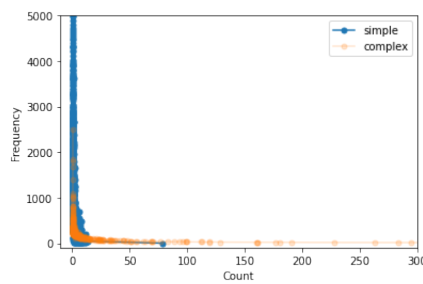
The cutoff frequencies and cutoff percentiles (if the words have frequencies below the percentile, then they are complex words) are shown in Table 6.2. The accuracy distributions across possible cutoff frequencies for the three datasets are shown in Figure 6.2. The classification accuracy reaches a peak very quickly for all three datasets: The frequency cutoff is 154, 21, and 168 respectively for the three datasets, and there exist a considerable number of words with frequencies in the hundreds and thousands (see Figure 6.2). Then, the accuracy slightly drops and hits a plateau, except in the case of the CLASSIFICATION dataset in which the accuracy remains very



(a) READABILITY



(b) CLUSTER



(c) CLASSIFICATION

Figure 6.1: The frequency distribution of the three full (unsplit) datasets.

high after beyond the peak accuracy point found from the training dataset.

We observe a considerable difference between the cutoff frequency for the CLUSTER dataset and that of the other two datasets. This is probably due to the fact that CLUSTER is a much smaller corpus: it has around a half million words compared to the other two datasets which respectively contain 1.5 million and approximately 5 million words. Thus, the classification process for CLUSTER is more likely to be affected by different noises. However, the cutoff percentiles are more uniform across the datasets, landing at around the 80-90% mark.

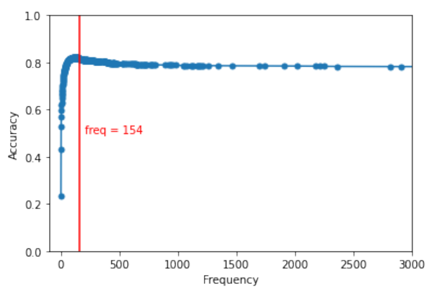
The accuracy, precision, recall and f-1 scores are reported in Table 6.1. All of these scores surpass 80%, which demonstrates a reliable performance of this method across the datasets. Recall scores are also high across the three datasets, with an average of 96.47%.

	accuracy	precision	recall	f-1
READABILITY	0.8174	0.9240	0.9717	0.9473
CLUSTER	0.8358	0.8103	0.9365	0.8689
CLASSIFICATION	0.9529	0.9351	0.9860	0.9599

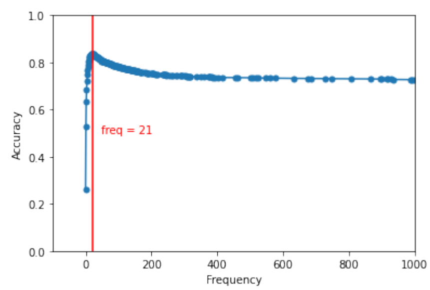
Table 6.1: The accuracy, precision, recall, and f-1 scores of the frequency cutoff approach across the three testing datasets

	cutoff frequency	cutoff percentile
READABILITY	154	0.9159
CLUSTER	21	0.7956
CLASSIFICATION	168	0.9255

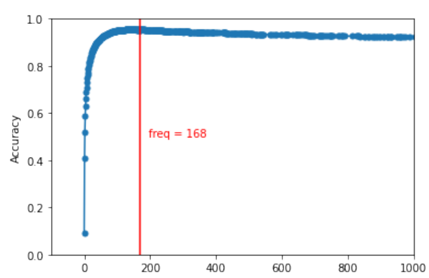
Table 6.2: The cutoff frequency and the cutoff percentile of the three testing datasets



(a) READABILITY



(b) CLUSTER



(c) CLASSIFICATION

Figure 6.2: The accuracy distributions across possible cutoff frequencies of the three testing datasets.

	accuracy	**precision**	**recall**	**f-1**
READABILITY	0.8207	0.8195	0.9831	0.8939
CLUSTER	0.8248	0.8212	0.9668	0.8880
CLASSIFICATION	0.9540	0.9581	0.9923	0.9750

Table 6.3: The accuracy, precision, recall, and f-1 scores of the SVM classifier of the three testing datasets

6.2 Support Vector Machines (SVM) Classifier

Support Vector Machines are used to classify the complex and simple words. The features used are: frequency, number of syllables, number of characters, and number of characters and diacritics. The regularization parameter C is 1 as we aim to classify all examples correctly. The kernel used for the experiments is the Radial Basis Function kernel:

$$k(x, z) = \exp(-\gamma||x - z||^2), \tag{6.1}$$

where γ is 1/number of features, which is 0.25.

The accuracy, precision, recall and f-1 are reported in Table 6.3. We observe that all of the scores also surpass 80% and the majority of them reaching the 90% mark, with recall exceeding 95% for all three of the datasets.

Chapter 7

Human Annotation

To quantify the quality of the datasets for the automated Complex Word Identification task in Vietnamese, three participants are asked to manually classify 100 words sampled from the Simple Word List and 99 words sampled from the READABILITY Complex Word List. All participants are native Vietnamese speakers pursuing a college degree in the United States. The instructions are provided in Vietnamese (shown in Figure 7.1), in which an example of one simple word and one complex word is demonstrated. The participants are reassured that there are no right or wrong answers, encouraged to use their intuition when making the decision, and 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 2 out of 3 participants. Then, we compute the accuracy, precision, recall and f-1 scores of the annotation process against our datasets. The results are reported in Table 7.1).

We observe 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%, mean-

	accuracy	**precision**	**recall**	**f-1**
All	0.4372	0.7273	0.4586	0.5625
Majority	0.8241	1.0	0.7388	0.8498

Table 7.1: The accuracy, precision, recall, and f-1 scores of the human annotation process

Bên cạnh mỗi từ
- điền 0 nếu thấy từ đó dễ (dễ hiểu, dễ phân biệt các nghĩa)
- điền 1 nếu thấy từ đó khó (khó hiểu, khó phân biệt các nghĩa)
- nếu phân vân, điền 1

LƯU Ý: Không có đáp án đúng hay sai.

VÍ DỤ

Word	Complex	
gia đình	0	Từ này với mình là một từ dễ vì được sử dụng hàng ngày
khải hoàn	1	Từ này với mình là một từ khó vì mình không sử dụng nó nhiều

(a) Vietnamese instructions

Next to each word
- write down 0 if you think the word is simple (easy to understand, easy to differentiate between different meanings)
- write down 1 if you think the word is complex (hard to understand, hard to differentiate between different meanings)
- if ambivalent, write down 1

NOTE: There is no right or wrong answer

EXAMPLE

Word	Complex	
gia đình (family)	0	This word is simple for me because I use it in my everyday life
khải hoàn (return - SV)	1	This word is hard for me because I don't use it often

(b) English instructions

Figure 7.1: The classification instructions for participants in Vietnamese and English [SV - Sino-Vietnamese words]

ing 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.

Chapter 8

Discussion

Frequency is an overpowering signal in determining whether a word is complex or simple as shown by the accuracy, precision, recall and f-1 scores of the **Frequency Threshold** experiment, which are all are greater than 0.8 (see Table 6.1). 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 results obtained from the Complex Word Classification task in English [PS16a].

The classification results for the CLASSIFICATION dataset are particularly impressive: all scores are greater than 0.93. This can be explained by the fact that there is less overlap between complex words and simple words in terms of frequency compared to the other two datasets (see Figure 6.1), so a horizontal line can be drawn to separate the two classes with minimal error.

By determining the frequency of the vocabulary in any corpus, based on our results (see Table 6.2), we can infer that the most frequent 10-20% of the words are simple, while the rest are complex.

These results across the three corpora are obtained even though there exist certain shortcomings in the datasets that may affect the performance. There exist words in the Simple 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 used to improve the classification results obtained from using a frequency threshold. 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 f-1 score (from 94.73% to 89.39%) of 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 labelling 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' labelling of complex words is consistent with the data we obtain, which can indicate the suitability of the READABILITY dataset for the CWI training purposes.

Chapter 9

Future Direction

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 counts (number of entries in a dictionary for example), synonym counts, and word type (whether the word is loan word).

Vietnamese Language Model: A Vietnamese language model can be used to provide contexts of the words that can improve classification performance. Several Vietnamese language models have been developed, such as PhoBERT [NN20], a pretrained language model that produces better results than the pretrained multilingual model XLM-R [CKG^{+}19] and contributes to the state-of-the-art performances of NLP tasks including Part-of-speech tagging, Dependency parsing, Named-entity recognition and Natural language inference.

Transfer Learning: Transfer learning can be used to apply the inferences learned for a high-resource language to Vietnamese, a low resource language. This method has been conducted for neural machine translation and shown to generate effective results under low-resource conditions, such as in [KB18].

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.

Next Steps in the Lexical Simplification pipeline: With the results obtained using a frequency threshold, attempt can be made at solving the Substitution Generation, Substitution Selection and Substitution Ranking step of the Lexical Simplification Pipeline. Considering the potential benefits of Text Simplification to a variety of target audiences, solving the automation question will introduce more helpful textual resources to different groups.

Chapter 10

Conclusion

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, 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 an average 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 surfave-level word features such as number of syllables and number of characters only marginally improve the recall scores but makes no significant difference in terms of accuracy, precision and f-1 scores. The Human Annotation experiment demonstrates how with a small number of annotators and a small sample, we can quantify how one dataset align 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.

Bibliography

[AG10] Sandra Aluísio and Caroline Gasperin. Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, 2010.

[Alv95] Mark Alves. Tonal features and the development of vietnamese tones. *Working Papers in Linguistics*, 27:1–13, 1995.

[Alv06] Mark Alves. Linguistic research on the origins of the vietnamese language: An overview. *Journal of Vietnamese Studies*, 1(1-2):104–130, 2006.

[Alv09] Mark J Alves. Loanwords in vietnamese. 617:637, 2009.

[AMSS20] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187, 2020.

[Anh21] Phan Anh. Vietnamese world's 21st most spoken language: Ranking - vnexpress international, Aug 2021.

[ATA21] Suha S Al-Thanyyan and Aqil M Azmi. Automated text simplification: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

[ATTQ15] Tran Ngoc Anh, Nguyen Phuong Thai, Dao Thanh Tinh, and Nguyen Hong Quan. 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, 2015.

[BN09] Marc Brysbaert and Boris New. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990, 2009.

[BRDS12] Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of COLING 2012*, pages 357–374, 2012.

[BSM12] Stefan Bott, Horacio Saggion, and Simon Mille. Text simplification tools for spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1665–1671, 2012.

[BT13] Gianni Barlacchi and Sara Tonelli. 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, 2013.

[Bur21] US Census Bureau. Top languages other than english spoken in 1980 and changes in relative rank, 1990-2010, Oct 2021.

[CDB96] Raman Chandrasekar, Christine Doran, and Srinivas Bangalore. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996.

[CK11] William Coster and David Kauchak. Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation*, pages 1–9, 2011.

[CKG+19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[CLMM07] Scott A Crossley, Max M Louwerse, Philip M McCarthy, and Danielle S McNamara. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30, 2007.

[CMC+98] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 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, 1998.

[CMKP13] Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. 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, 2013.

[Dev98] Siobhan Devlin. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*, pages 161–173, 1998.

[Dif89] Gérard Diffloth. Proto-austroasiatic creaky voice. *Mon-Khmer Studies*, 15:139–154, 1989.

[ĐDL98] Hữu Đạt, Trần Trí Dõi, and Đào Thanh Lan. Cơ sở tiếng việt (basis of vietnamese), 1998.

[DLN+08] Quang Thang Dinh, Hong Phuong Le, Thi Minh Huyen Nguyen, Cam Tu Nguyen, Mathias Rossignol, and Xuan Luong Vu. Word segmentation of vietnamese texts: a comparison of approaches. In *6th international conference on Language Resources and Evaluation-LREC 2008*, 2008.

[Doa99] TT Doan. Ngữ âm tiếng việt. *Vietnamese Phonetics), Hanoi National University Publishing House*, pages 99–148, 1999.

[DU06] Siobhan Devlin and Gary Unthank. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 225–226, 2006.

[EDS12] Soojeong Eom, Markus Dickinson, and Rebecca Sachs. Sense-specific lexical information for reading assistance. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 316–325, 2012.

[EOD14] Richard Evans, Constantin Orasan, and Iustin Dornescu. An evaluation of syntactic simplification rules for people with autism. Association for Computational Linguistics, 2014.

[Eva11] Richard J Evans. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and linguistic computing*, 26(4):371–388, 2011.

[FGWF14] Thomas François, Núria Gala, Patrick Watrin, and Cédrick Fairon. Flelex: a graded lexical resource for french foreign learners. In *International conference on Language Resources and Evaluation (LREC 2014)*, 2014.

[GŠ15] Goran Glavaš and Sanja Štajner. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, 2015.

[HCN00] Marcella Hu Hsueh-Chao and Paul Nation. Unknown vocabulary density and reading comprehension. 2000.

[HdGS+17] Eva Hasler, Adrià de Gispert, Felix Stahlberg, Aurelien Waite, and Bill Byrne. Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 45:221–235, 2017.

[HDLNN07] Vu Cong Duy Hoang, Dien Dinh, Nguyen Le Nguyen, and Hung Quoc Ngo. 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, 2007.

[HLMS12] Bui Thanh Hung, Nguyen Le Minh, and Akira Shimazu. Sentence splitting for vietnamese-english machine translation. In *2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 156–160. IEEE, 2012.

[Hoa17] Phe Hoang. *Từ điển Tiếng Việt (Vietnamese Dictionary)*. Da Nang Publising House, 2017.

[HS10] Michael Heilman and Noah A Smith. 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, 2010.

[H0] Cao Xuân Hạo. Tiếng việt-mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa (vietnamese-some questions on phonetics, syntax and semantics). *NXB Giáo dục, Hanoi*, 2000.

[KB18] Tom Kocmi and Ondřej Bojar. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*, 2018.

[KCZT10] Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, volume 2010, page 366. American Medical Informatics Association, 2010.

[Kes12] Robin Keskisärkkä. Automatic text simplification via synonym replacement, 2012.

[KKM04] Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. Text simplification for information-seeking applications. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 735–747. Springer, 2004.

[Kon16] Michal Konkol. Uwb at semeval-2016 task 11: Exploring features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1038–1041, 2016.

[KY19] Akihiro Katsuta and Kazuhide Yamamoto. Improving text simplification by corpus expansion with unsupervised learning. In *2019 International Conference on Asian Language Processing (IALP)*, pages 216–221. IEEE, 2019.

[Lau89] Batia Laufer. What percentage of text-lexis is essential for comprehension. *Special language: From humans thinking to thinking machines*, 316323, 1989.

[LEK+13] Gondy Leroy, James E Endicott, David Kauchak, Obay Mouradi, and Melissa Just. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of medical Internet research*, 15(7):e144, 2013.

[LLWL13] Chen Li, Fei Liu, Fuliang Weng, and Yang Liu. Document summarization via guided sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 490–500, 2013.

[LND18] An-Vinh Luong, Diep Nguyen, and Dien Dinh. A new formula for vietnamese text readability assessment. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 198–202. IEEE, 2018.

[LND20] An-Vinh Luong, Diep Nguyen, and Dien Dinh. Building a corpus for vietnamese text readability assessment in the literature domain. *Universal Journal of Educational Research*, 8(10):4996–5004, 2020.

[Mas12] Henri Maspero. Etudes sur la phonétique historique de la langue annamite. les initiales. *Bulletin de l'École française d'Extrême-Orient*, 12(1):1–124, 1912.

[Mas78] Jana M Mason. Facilitating reading comprehension through text structure manipulation. *Center for the Study of Reading Technical Report; no. 092*, 1978.

[MDZ16] Shervin Malmasi, Mark Dras, and Marcos Zampieri. Ltg at semeval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 996–1000, 2016.

[MHM+19] Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Antoine Bordes, Éric Villemonte de La Clergerie, and Benoît Sagot. Reference-less quality estimation of text simplification systems. *arXiv preprint arXiv:1901.10746*, 2019.

[MNS14] Angrosh Annayappan Mandya, Tadashi Nomoto, and Advaith Siddharthan. Lexico-syntactic text simplification and compression with typed dependencies. In *25th International Conference on Computational Linguistics*, 2014.

[MSMT10] Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. Entity-focused sentence simplification for relation extraction. In *Proceedings of the 23rd International Conference* *on Computational Linguistics (Coling 2010)*, pages 788–796, 2010.

[MY17] Takumi Maruyama and Kazuhide Yamamoto. Sentence simplification with core vocabulary. In *2017 International Conference on Asian Language Processing (IALP)*, pages 363–366. IEEE, 2017.

[NGCS17] Shashi Narayan, Claire Gardent, Shay B Cohen, and Anastasia Shimorina. Split and rephrase. *arXiv preprint arXiv:1707.06971*, 2017.

[Ngu97] H Nguyen. London oriental and african language library: Vietnamese, 1997.

[NN20] Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*, 2020.

[NNLNH06] Cam-Tu Nguyen, Xuan-Hieu Nguyen, Trung-Kien a nd Phan, Minh Le Nguyen, and Quang Thuy Ha. 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, 2006.

[NNN+18] Minh-Tien Nguyen, Hoang-Diep Nguyen, Van-Hau Nguyen, et al. Towards state-of-the-art baselines for vietnamese multi-document summarization. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 85–90. IEEE, 2018.

[NNV+17] Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. A fast and accurate vietnamese word segmenter. *arXiv preprint arXiv:1709.06307*, 2017.

[NŠPD17] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 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, 2017.

[NTNN06] Thanh V Nguyen, Hoang K Tran, Thanh TT Nguyen, and Hung Nguyen. Word segmentation for vietnamese text categorization: an online corpus approach. *RIVF06*, 2006.

[ODL+07] Ethel Ong, Jerwin Damay, Gerard Lojico, Kimberly Lu, and Dex Tarantan. Simplifying text in medical literature. *Journal of Research in Science, Computing and Engineering*, 4(1):37–47, 2007.

[OEM18] Constantin Orăsan, Richard Evans, and Ruslan Mitkov. Intelligent text processing to help readers with autism. In *Intelligent Natural Language Processing: Trends and Applications*, pages 713–740. Springer, 2018.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[PS13] Gustavo Paetzold and Lucia Specia. Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 116–125, 2013.

[PS16a] Gustavo Paetzold and Lucia Specia. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, 2016.

[PS16b] Gustavo Paetzold and Lucia Specia. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[PS17] Gustavo H Paetzold and Lucia Specia. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593, 2017.

[PTMHR+08] Hông Phuong, Nguyên Thi Minh Huyên, Azim Roussanaly, Hô Tuòng Vinh, et al. A hybrid approach to word segmentation of vietnamese texts. In *International conference on language and automata theory and applications*, pages 240–249. Springer, 2008.

[PVG+11] 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Q+77] Stephen P Quigley et al. The language structure of deaf children. *Volta Review*, 79(2):73–84, 1977.

[QKCH17] Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *Journal of medical Internet research*, 19(12):e417, 2017.

[RBYDMS13] Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer, 2013.

[Rob81] Nancy Lee Robbins. *The effects of signed text on the reading comprehension of hearing impaired children*. PhD thesis, The University of Nebraska-Lincoln, 1981.

[Rud93] Alan P Rudell. Frequency of word usage and perceived word difficulty: Ratings of kučera and francis words. *Behavior Research Methods, Instruments, & Computers*, 25(4):455–463, 1993.

[Sag18] Horacio Saggion. Text simplification. In *The Oxford Handbook of Computational Linguistics 2nd edition*. 2018.

[SAR18] Elior Sulem, Omri Abend, and Ari Rappoport. Semantic structural evaluation for text simplification. *arXiv preprint arXiv:1810.05022*, 2018.

[SB12] Sara Botelho Silveira and António Branco. Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pages 482–489. IEEE, 2012.

[SBM17] Isabel Segura-Bedmar and Paloma Martínez. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of biomedical semantics*, 8(1):1–9, 2017.

[SFS+17] Horacio Saggion, Daniel Ferrés, Leen Sevens, Ineke Schuurman, Marta Ripollés, and Olga Rodríguez. Able to read my mail: An accessible e-mail client with assistive technology. In *Proceedings of the 14th International Web for All Conference*, pages 1–4, 2017.

[ŠFSRP18] Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. Cats: A tool for customized alignment of text simplification corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[Sha13] Matthew Shardlow. 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, 2013.

[Sha14a] Matthew Shardlow. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *LREC*, pages 1583–1590, 2014.

[Sha14b] Matthew Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70, 2014.

[She85] Cynthia M Shewan. Auditory comprehension problems in adult aphasic individuals. *Human Communication Canada*, 9(5):151–155, 1985.

[SJG11] Norbert Schmitt, Xiangying Jiang, and William Grabe. The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1):26–43, 2011.

[SM14] Advaith Siddharthan and Angrosh Mandya. 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, 2014.

[SM20] Punardeep Sikka and Vijay Mago. A survey on text simplification. *arXiv preprint arXiv:2008.08612*, 2020.

[SNM04] Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. Syntactic simplification for improving content selection in multi-document summarization. 2004.

[SŠB+15] Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36, 2015.

[Sta20] Felix Stahlberg. Neural machine translation: A review and survey. 2020.

[Tab38] Jean-Louis Tabard. Introduction to: Pierre-joseph pigneau and jean-louis tabard, dictionarium anamitico-latinum. *Serampore: J. Marshman*, 1838.

[Tan07] Giang Tang. Cross-linguistic analysis of vietnamese and english with implications for vietnamese language acquisition and maintenance in the united states. *Journal of Southeast Asian American Education and Advancement*, 2(1):3, 2007.

[TNN+20] Nhi-Thao Tran, Minh-Quoc Nghiem, Nhung TH Nguyen, Ngan Luu-Thuy Nguyen, Nam Van Chi, and Dien Dinh. Vims: a high-quality vietnamese dataset for abstractive multi-document summarization. *Language Resources and Evaluation*, 54(4):893–920, 2020.

[VD01] George Van Driem. *Languages of the Himalayas: an ethnolinguistic handbook of the greater Himalayan region*, volume 2. Brill, 2001.

[VK08] David Vickrey and Daphne Koller. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*, pages 344–352, 2008.

[VNN+18] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. Vncorenlp: A vietnamese natural language processing toolkit. *arXiv preprint arXiv:1801.01331*, 2018.

[VSBN07] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618, 2007.

[WJU+09] Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. Facilita: reading assistance for low-literacy readers. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36, 2009.

[WL11a] Kristian Woodsend and Mirella Lapata. 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, 2011.

[WL11b] Kristian Woodsend and Mirella Lapata. Wikisimple: Automatic simplification of wikipedia articles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 927–932, 2011.

[XCBN15] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

[XNP+16] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.