5-2019

Automatic Scaling of Text for Training Second Language Reading Comprehension

Justin Vasselli

Automatic Scaling of Text for Training Second Language Reading Comprehension

by

Justin Vasselli

FINAL PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS IN LIBERAL STUDIES

SKIDMORE COLLEGE
May 2019
Advisors: Andrew Cencini, Hope Casto

THE MASTER OF ARTS PROGRAM IN LIBERAL STUDIES
SKIDMORE COLLEGE

Table of Contents

List of Tables:

List of Figures:

List of Images:

Abstract

For children learning their first language, reading is one of the most effective ways to acquire new vocabulary. Studies link students who read more with larger and more complex vocabularies. For second language learners, there is a substantial barrier to reading. Even the books written for early first language readers assume a base vocabulary of nearly 7000 word families and a nuanced understanding of grammar. This project will look at ways that technology can help second language learners overcome this high barrier to entry, and the effectiveness of learning through reading for adults acquiring a foreign language. Through the implementation of Dokusha, an automatic graded reader generator for Japanese, this project will explore how advancements in natural language processing can be used to automatically simplify text for extensive reading in Japanese as a foreign language.

Introduction

Reading is a complex cognitive process that improves with practice and exposure to the correct difficulty level of text, and results in a myriad of benefits to language skills. Because reading is such a natural and integral component of language acquisition, most native speakers of a language advance their lexicon through reading in their native language as soon as they are capable of it. While just as beneficial for foreign language study, reading in a second language is a much more daunting task. To become proficient readers, second-language learners have to work around their linguistic disadvantages. The correct level of text for a native speaker, who has years of acquiring language orally as a solid linguistic foundation for literacy, is not an ideal entry point for the second language learner, who generally studies the written and oral aspects of their target language simultaneously. Still, there is good evidence in the literature that learning to read in a second language can create higher motivation for learning the language (Hitosugi & Day, 2004; Mason & Krashen, 1997; Swaffer & Woodruff, 1978), help learners build a more diverse and rich vocabulary (Horst, 2005; Nation, 2013; Pitts, White & Krashen, 1989), and can be a largely autonomous way to improve language skills (Benson, 2011).

In addition to its lexical benefits, reading has been shown to have a positive effect on learner motivation. When the University of Texas changed the focus of their German program in 1975 to deemphasize grammar and in-class language production in favor of learning through input, both verbal and written, they found the percent of students who elected to continue their German studies after the first term jumped from an average of 55% per year to 72% (Swaffer & Woodruff, 1978). The students in these classes were encouraged but not required to speak German during class. From the fifth week of the first term through

the remainder of the program, the focus of the course was primarily on reading comprehension. The students started by reading dialogues and progressed to newspapers by the end of the second semester class. In addition to the lower attrition rates, the students responded positively to the new curriculum in feedback forms, and scored higher than average on a national test conducted by the Modern Language Association. Other researchers have also found that a curriculum structured around reading contributes significantly toward increasing learner motivation (Hitosugi & Day, 2004; Mason & Krashen, 1997).

Second language (L2) vocabulary can be acquired through the reading of large amounts of comprehensible text as demonstrated by Saragi, Nation, and Meister (1987). In this study, which examined how vocabulary is acquired from comprehensible context, students were asked to read *A Clockwork Orange* in their native English. The students were subsequently tested on their understanding of 90 of the Russian slang words that appeared in the book. The average score on the vocabulary test was 76%, with higher scores on a word being correlated with the number of times it appeared in the book. This suggested that given enough exposure to a word in context, the word will be acquired by the reader. This study inspired a number of other studies on L2 vocabulary acquisition through extensive reading of both authentic texts (Horst, Cobb, & Meara, 1998; Pellicer-Sánchez & Schmitt, 2010; Pitts et al., 1989) and graded readers -- books written specifically for beginning second language learners (Horst, 2005; Waring & Takaki, 2003).

Reading also has a powerful role to play in non-classroom study. One of the appealing points of studying a language through reading is that reading can be a largely autonomous and private way of improving language skills in general. Pickard (1995)

released a case-study following three proficient German students studying English at the University of Humberside. Through interviews and questionnaires, Pickard found that these successful students used a wide array of strategies to advance their language skills outside of the classroom. The students in the study reported that they believed reading newspapers and novels was an extremely important contributing factor to their success at learning English. This study highlights one of the most compelling reasons for language learners to learn to read: it is a largely self-directed activity. While teachers can certainly scaffold and help teach literacy, reading is not dependent on having access to a speaker of the language in the way that conversation and communication based approaches are. This may be why reading is a method that autonomous learners commonly employ when learning a second language (Eskey, 2002; Hyland, 2004; Leung, 2002, Senoo & Yonemoto, 2014).

Learning to read in a foreign language strengthens language skills and is enjoyable enough to help learners stay interested and motivated over time. Linguistic benefits aside, reading is also a very accessible activity. It is available as a way to build students' language comprehension both in and out of the classroom, as well as that of independent learners regardless of their access to native speakers. Texts are available in almost any language online and so, unlike conversation, which requires a speaker to converse with, access to text is easy for most people. Perhaps due to the comparative difficulty of studying independent language learners, there is a dearth of studies on independent learners in general, and most of the studies on the benefits of reading have been conducted on students in classroom settings. Despite this, reading can be used to great effect outside of traditional studies as long as the appropriate text is available (Senoo & Yonemoto, 2014).

Yet finding level-appropriate texts for less popular second languages, such as Japanese, can be incredibly difficult for classroom and independent students alike. Because of this lack of leveled reading resources, students may not have access to enough sufficiently motivating, appropriately-leveled text for reading to be an effective learning tool for them (Hitosugi & Day, 2004). Given the crucial role reading plays in the L2 acquisition process, lack of adequate reading material to meet a student's needs at their unique level could present a significant obstacle to learning (Day & Bamford, 1998; Elley & Mangubhai, 1983; Krashen, 1985). Here, recent developments in the field of computer science known as natural language processing (NLP) could play a role in solving the problem of creating appropriately-leveled language learner text.

If a computer program could be constructed to reduce the level of a piece of text automatically, then it would eliminate the problem of finding a large enough variety of appropriately leveled text for a Japanese extensive reading program. Additionally, independent learners would be able to self-convert text found online to the ideal level for them to read fluently for meaning. It would even be possible for the same text to be scaled to multiple levels, enabling learners to work their way into more complicated language with a solid understanding of the content of the text obtained through initial readings at lower difficulty levels, or for students at different reading levels in a group class to work through the same text at the level most appropriate to their current L2 learning needs. To demonstrate this, the design, implementation, and analysis of the automatic graded text generator Dokusha will be examined.

Children's Literature in the L2 Classroom

Making the appropriate level of text available to students is not always an easy task. Extensive reading programs in English classrooms typically give students access to a library of graded readers, texts written at a specific level of language for foreign language learners. This allows students to select books that interest them at their current level. However, graded readers in non-English languages are not as common, and the smaller selection prevents learners from accessing texts that they are motivated to read at the correct difficulty level for them. As a result of this lack of appropriate language learner literature, extensive reading programs in other languages may resort to using children's literature (Hitosugi & Day, 2004).

Children's literature is not an ideal substitute for graded readers. While written for children without the full lexicon of an adult native speaker, children's literature still assumes a much more complete understanding of vocabulary and an implicit understanding of the complex grammar of the language. Although Day and Bamford (1998) claim that "If a language lacks language learner literature, teachers can turn to a sure source of easy reading material that exists in almost every language: books designed to teach children their first language" (p 98, as cited by Webb & Macalister, 2013), there are several reasons why children's literature is not ideal for L2 readers. First, the vocabulary that shows up in children's literature is more reflective of children's language, and could be very age specific (Webb & Rodgers, 2009). Children's language is often different from adult language and children's books may not model language as adults want to acquire it. For example, Japanese picture books aimed at 0-2 year olds are filled with onomatopoeia, and contain little sentence structure. These onomatopoeia may be useful for an advanced learner to grow their vocabulary, but won't serve a beginning student well in conversation. Additionally, the

vocabulary load of children's books aimed beyond the 0-2 year old age range is often too intense for beginners to read fluently. It is also not easy to judge how appropriate the level of a book is for a given learner, as an L2 learner's language is very different from the developing native language of a young child (Hayes & Ahrens, 1988; Hirsh & Nation, 1992; Webb & Macalister, 2013).

An additional problem with utilizing children's literature as a means to teach literacy to adult second language learners is that books targeted at children may not be motivating for adult learners to read. Since one of the most important motivating factors for students in extensive reading programs is having access to text they want to read (Farrell, 2009), this makes children's literature a less than ideal material for extensive reading programs.

Another surprising feature of children's literature may present an even greater obstacle for L2 readers: the language of children's books, while ostensibly simpler than texts aimed at adults, nevertheless requires a surprisingly large lexicon to process efficiently. Webb and Macalister (2013) found that texts aimed at 7-13 year olds contained nearly the same vocabulary requirement as text written for ages 12 and up. In order to reach 98% coverage in a text, the amount needed for fluent reading, readers of authentic children's literature needed to have a vocabulary size of 10,000 English word families, compared to the 3,000 word families required for the same level of comprehension of graded readers.

Hitosugi and Day (2004) conducted one of the only studies on extensive reading in the Japanese language classroom using children's literature. While their class size was not large enough for definitive results, they found that, in general, more students reported enjoying the class that featured extensive reading than the students in the traditional classes, which feature reading comprehension passages and explicit grammar instruction. Students

who engaged in extensive reading scored better on a traditional reading comprehension assessment than students in the control classes, despite not having short-passage reading and comprehension questions as a component of the extensive reading course. Hitosugi and Day's extensive reading course appears to have met the goals of the program, but with other researchers concluding that children's literature is not at an appropriate level for beginner L2 readers (Webb & Macalister, 2013), it is a reasonable hypothesis that using graded readers could prove even more effective in a Japanese-language extensive reading program, and this is an area ripe for further research. For this to be possible, more graded readers are needed for Japanese language learners.

Extensive Reading

Extensive reading is a term used to distinguish reading a lot of text relatively quickly, from intensive reading of a text, which is often a word by word examination of a short passage. Extensive reading is typified by reading large quantities of text at a fast pace with a 'big picture' focus. This approach prioritizes comprehension of general meaning instead of careful analysis of how the language is used by the author (Farrell, 2009). This is a departure from the historically more predominant method of intensive reading, in which readings in foreign language classrooms have taken the form of short passages around 500-1000 words in length (Light, 1970). These short passages are typically read over the course of a week or more in the classroom, with careful attention paid to each unfamiliar vocabulary item and grammar point. Despite the goal of reading-comprehension passages being ostensibly for the students to improve their reading comprehension skills in a foreign language, the slow, methodical nature of intensive reading uses up too many cognitive resources to allow

students to derive meaning from the passage effectively (Nation, 2013). By contrast, extensive reading, where there is no careful analysis of vocabulary or structure, has been shown to increase reading speed and comprehension better than deep reading of short passages (Bell, 2001).

A delicate balance between new and learned vocabulary must be maintained in order to keep material at the ideal level for extensive reading. Nearly all of the words encountered in the text must be known by the students in order to build reading fluency (Farrell, 2009; Nation, 2013). Day and Bamford (2002) say that "for extensive reading to be possible and for it to have the desired results, text must be well within the learners' reading competence in the foreign language" (p. 137). Graded readers were created to address this problem. Graded readers have limited vocabulary; for example, graded readers for English learners often use only the most common 1000-3000 English word families. The targeted nature of graded readers lightens the language processing demands on the learner, allowing for a more fluent reading experience (Nation, 2013).

The success of book flood programs, in which classrooms in developing countries are given large quantities of books of high interest to the students through which to teach English, provide another compelling case for extensive reading programs (Elley, 2000). The essential point of a book flood program is that the books are not supplementing a traditional textbook approach. Instead, they provide the main means through which English is acquired by the students. The results of these programs show significant progress in reading skills by the students in the book flood participant schools as compared with the control schools. Book flood participant school reading gains doubled over that of control groups, and the

number of students who passed a standard English test from the experimental group was twice that of the control group (Elley, 2000).

Reading Comprehension

When altering authentic text for second language learners, it is important to understand the differences in process and need between first language readers and second language readers. Readers learning first versus second languages have very different needs (Grabe & Stoller, 2013; Koda, 2007; Perfetti & Dunlap, 2008); reading provides them with different goals and challenges, and they require different resources and solutions accordingly. Because of these differences, successful literacy teaching strategies for L1 versus L2 readers may look very different (Farrell, 2009; Lightbown & Spada, 2013).

Foreign language students often learn to read in their L2 at the same time that they are learning it orally, perhaps even using reading exercises starting from their very earliest stages of language learning. In contrast, by the time L1 students learn to read they have already tacitly acquired most of the basic grammatical structures of their native language (Grabe & Stoller, 2013). This means that L1 readers have a huge head start on grammar as well as vocabulary when compared with L2 readers.

To better understand these differing needs and how they might be addressed, it is important to understand what goes into the act of reading on a cognitive level. The activity of reading is a complex cognitive process that involves a combination of top-down and bottom-up processes (Farrell, 2009). Top-down processes are those that activate the reader's prior knowledge to influence how they interpret the text. Bottom-up processing relies first on decoding skills such as character and word recognition before organizing the information

based on the sentence syntax. It is generally thought that during reading these processes occur simultaneously and interact with each other (Sato, Matsunuma, & Suzuki, 2013; Stanovich, 1980). In order for the bottom-up process to work effectively, the reader must be able to process phonology (pronunciation), semantics (vocabulary), and syntax (grammar), and must have a sufficient working memory to store all the component pieces before they have been automatized. For the top-down process to work, the reader must access their prior knowledge and take cues from discourse markers, the words or phrases that provide structure to the text as a whole.

Beginning language learners usually use more of the bottom-up processes to extract meaning from the text until they have reached a high enough level of lexical knowledge to process the text efficiently using a combination of top-down and bottom-up approaches. This has been shown to be as true for Japanese learners (Everson & Kuriya, 1999; Horiba, 1990) as well as English learners (Lee & Schallert, 1997; Yamashita, 2002). Native language reading strategies cannot carry over to the target language until a certain level of fluency is reached, allowing the text to be decoded automatically (Clarke, 1980; Cummins, 1979).

The different background knowledge and language learning trajectories of L1 versus L2 readers explains why teaching strategies developed for L1 readers have been shown to be less effective on L2 readers. The implications of these differences in learning process are significant; second language learners differ from first language learners in the breadth of their vocabulary, their ability to process syntax, and even the way they visually process the characters they are reading (Clarke,1980; Grabe & Stoller, 2013; Koda, 2008).

Lexical Processing

Readers process new or low frequency words through the following neural path: the orthographic processor triggers the phonological processor, which may need to gather information from a contextual processor before processing the semantics of the word (Muljani, Koda, & Moates, 1998). Through exposure to enough text, this process becomes more automatic, allowing for a word to go directly from the orthographic processor to the semantic processor. A word that has become automatized in this way is referred to as within the reader's "sight vocabulary" (Mezynski, 1983). Sight vocabulary typically requires little cognitive processing, allowing for more cognitive resources to be used in processing the language at a higher level, such as comprehending the text as a whole and understanding the implications of the language used (Mezynski, 1983).

The words in a piece of text that are within a reader's sight vocabulary are referred to as the lexical coverage of that text (Nation, 2013; Webb & Macalister, 2013). Lexical coverage is typically expressed as a percentage of the running words in the text, where repeated words are counted multiple times towards the percentage calculation. If an encountered word is not within the sight vocabulary of the reader, they will try to guess from the surrounding context. This guessing process taxes the working memory of the reader, slowing down the reading process and making it difficult for them to construct the meaning of the full sentence. Thus, the lexical coverage of a text with respect to a given reader is an essential element of its overall readability.

Laufer and Sim (1985) found that when reading a difficult passage, L2 readers will try to understand the words first, ignoring the words and phrases they can't guess, and sometimes mistaking unfamiliar words for familiar ones that are visually similar. Without a

reasonably complete understanding of the vocabulary in a sentence, the meaning of the sentence is lost for the reader. This finding has led researchers to focus on finding the lexical threshold for reading comprehension; that is, how many lexical items outside the reader's sight vocabulary a text can contain before it becomes incomprehensible to the reader (Cobb, 2007; Hu & Nation, 2000; Laufer, 1989; Nation, 2006).

In order to achieve "adequate comprehension" of an English text when reading without the assistance of a dictionary, it is generally accepted that a lexical coverage of sight vocabulary of 98% is required. This coverage requires that only one in every twenty words be guessed from context. This number was reached by several researchers. Laufer (1989) found there to be significantly more students who passed a reading comprehension test with a lexical coverage of 95% than the students with smaller vocabularies. In addition, Hu and Nation (2000) used regression analysis to conclude that a coverage of 98% would be a reasonable threshold for most texts with relatively simple grammar, and that coverage should be higher for newspapers and academic text.

Adequate vocabulary is considered to be the most influential factor in determining whether or not a text is comprehensible. Clarke (1980) found that poor readers tend to use grammatical cues over lexical cues in both their first and second languages. By comparison, good readers use more lexical cues than grammatical cues in their first language. This appears to support the hypothesis that the most important component of reading comprehension is an adequate vocabulary.

Syntactic Processing

The contributions of vocabulary knowledge to reading comprehension have been well studied, but the correlation between syntactic processing, or grammar knowledge, and reading comprehension skills has been researched much less in both L1 and L2 settings. This may be due to the difficulty in reliably testing grammar skills as distinct from reading skills. Since typical grammar skill measures involve reading full sentences, many studies that attempt to calculate the contribution of grammar knowledge to reading comprehension have been criticized for failing to appropriately isolate the two skills (Shiotsu & Weir, 2007). It is clear that grammar plays some part in reading comprehension, but it remains unclear how much it contributes (Grabe & Stoller, 2013).

Syntactic processing skills are broken into two categories: syntactic awareness and syntactic knowledge. Syntactic awareness is the metalinguistic skill of being able to reflect on or manipulate grammar, and syntactic knowledge is defined as the ability of a person to comprehend or produce grammatical phrases (Brimo, Apel, & Fountain, 2017). Language classrooms or textbooks that focus on grammar and form serve to build syntactic awareness, but may not teach syntactic knowledge as effectively. Most native speakers of a language have a higher degree of syntactic knowledge than syntactic awareness. They have an implicit understanding of how words fit together, but may not be able to articulate the underlying rules. This is unfortunate because syntactic awareness has not been found to contribute directly to reading comprehension (Brimo et al., 2017). On the other hand, several studies have found that syntactic knowledge, the more implicit linguistic skill, contributes significantly to reading comprehension (Brimo et al., 2017; Shiotsu & Weir, 2007).

The ways in which syntactic processing contributes to reading comprehension can be

clearly observed when looking at the processing of complex sentences. As the goal of this project is to alter Japanese text in particular, it is important to keep in mind language specific barriers to syntactic comprehension. In one of the few studies on Japanese syntactic processing published in English, Muto (2015) found that complex sentences place a greater burden on working memory and verbal comprehension skills than linear sentences. One type of complex sentence is a garden path sentence, defined as a sentence in which the reader has to reinterpret their built-up construct of the sentence mid-reading. For example, the English sentence "Put the frog on the napkin into the box" leads the reader to initially believe the directive is to place the frog on the napkin, until the word ‘into’, when the reader reinterprets the sentence to mean that the frog is already on the napkin, and should be moved from the napkin to the box (Pozzan & Trueswell, 2016).

The garden path sentences that Muto (2015) studied were in Japanese. In Japanese, a clause can be used to modify a noun in such a way that the clause reads as a complete sentence before the reader realizes that it is only a modifier. A Japanese example of a garden path sentence is generally constructed "S1 O2 V2 S2 V1", where the number indicates which predicate the subject/object belongs to. For example:

太郎	は	りんご	を	食べた	花子	と	会う
Tarō	*wa*	*ringo*	*wo*	*tabeta*	*Hanako*	*to*	*au*
Taro	TOPIC	apple	OBJECT	ate	Hanako	with	meet

"Taro will meet with Hanako, who has eaten an apple"

In the above sentence, a reader would assume Taro was the eater of the apple until encountering the name Hanako, as *Tarō wa ringo wo tabeta* translates to “Taro ate the apple”.

Then the reader would have to reevaluate the sentence to determine that Hanako is the one who ate the apple, and that Taro will meet her. These types of sentences seem to require more processing skills than simple sentences such as "Taro will meet with Hanako" or "Hanako has eaten an apple" (Muto, 2015). As the comprehension of garden path sentences has been shown to tax working memory in Japanese, these types of sentences should be removed at texts aimed at lower level Japanese readers in order to to construct a fluent reading experience for a language learner.

Visual Processing of Japanese

For the purpose of constructing Japanese texts for L2 readers, it is important to take into account the unique challenges of visually processing Japanese orthography (Yokoyama & Imai, 1989). As mentioned above, Japanese is typically written with a combination of character sets, an intermingling of native syllabaries and imported logograms. Written Japanese consists of three distinct character sets: kanji, hiragana, and katakana. Typically, compound nouns and the roots of verbs and adjectives are written in kanji, the Sino-Japanese logographic system. Function words, auxiliary verbs, and grammatical morphemes are nearly always written in the phonetic hiragana alphabet. Katakana, also a phonetic alphabet, is used mostly for Japanese's many 'loan' words (Sakuma, Sasanuma, Tatsumi, & Masaki, 1998), such as パン *pan*, meaning 'bread', which was taken from Portuguese.

Japanese orthography is fairly fluid in practice. It is not uncommon to see words that can be represented using kanji written out in hiragana, or katakana used stylistically to spell out words typically seen in hiragana or kanji. It is common in literature geared toward student learners of Japanese to write out kanji words in hiragana to reduce the number of

unfamiliar characters. This ostensibly makes reading easier for language learners, but this method may not actually be an effective mode of simplification, and may in fact slow the learning process over the long term rather than serving to accelerate it.

Studies of native Japanese speakers have found that native readers take longer to process words when they are presented in an atypical orthography (Yamada, Imai, & Ikebe, 1989). Umemura (1981) found that words written in hiragana were easier to recall on an immediate post-test, while words written in kanji were easier to recall on a delayed post test, indicating that words written as they traditionally are, in kanji, might be retained better long-term. Hiragana characters are phonetic symbols, while kanji are logograms. The researchers concluded that the immediate post test relied more on phonetic information stored in short term memory, and the delayed post test relied more on stored meaning information.

This finding raises an interesting question: is it more valuable for readers to process the phonetic information or semantic information from a written character? Kaiho (1975) found that kanji can be processed as meaning directly, without being converted into phonetic information, and that this allows the encoding in memory to able to be remembered more easily long-term.

Text that is written entirely in hiragana or katakana, collectively known as “kana”, is actually more difficult to process in some ways, in part because Japanese doesn't use spaces between words. Kanji characters help visually segment sentences into words, making sentences with limited kanji more difficult for native and non-native readers of Japanese alike (Koda, 1992). Harada (1988) found that, in addition to native readers, non-native readers of Japanese also have a difficult time understanding text that is written entirely in kana. Harada compared reading comprehension and reading speed between three levels of non-native

readers and native readers on text presented three ways: authentic use of kanji, kana only with no spaces, and kana with spaces. The researcher found that even for the lowest level of non-native readers tested, reading comprehension was higher when kanji was provided. The native readers had no comprehension difficulties regardless of orthography, but the time it took them to read was much greater with the all-kana text. It was hypothesised that non-native readers use kanji primarily to discern word boundaries, but for the non-native readers tested, the addition of spaces could not fully compensate for the loss of information the kanji themselves provide (Harada, 1988). Interestingly, the most advanced group of non-native readers had more trouble interpreting the kana text without spaces than any other group.

The particular difficulty of non-native readers understanding kana-only text highlights the need for texts aimed at L2 readers, rather than using children's books for study. Japanese children's books, especially picture books, are written nearly entirely in kana. This allows such books to be more accessible to Japanese children, who have a large oral vocabulary, but haven't been exposed to many kanji. L2 readers should be learning kanji as they learn new vocabulary, and should be exposed to the kanji often to acquire automaticity in reading the characters.

Words typically written in kanji are jarringly unfamiliar when written in kana. A native English speaker may experience a similar unfamiliar feeling when encountering text with spelling errors in English. Consider the sentence "Zombies eat branes for breakfast." The word "brane" is phonologically the same as "brain" in English, but orthographically different, what Besner (1990) refers to as a pseudohomophone, and would be confusing to encounter in an English text where a reader would be expecting "brain". If a native Japanese reader is unlikely to encounter a given word written in kana, offering it to a language learner

in kana is doing them a disservice, because it is acclimating them to an orthographically atypical form. Instead, it would be better to give language learners sufficient exposure to the kanji form, in order to increase familiarity with standard Japanese orthography (Kirwan, 2003).

A compromise must be found between presenting text with authentic orthography, which may be difficult for a learner to pronounce, and presenting text with only the phonetic scripts, which a learner is guaranteed to be able to pronounce, but may not understand as easily. This compromise often comes in the form of furigana, small phonetic pronunciations of kanji written above or below the logographic characters in horizontal script, or to the right of kanji in vertical script. Furigana enables the text to remain pronounceable, even with liberal use of kanji forms.

Learning Orthography

When learning an entirely new character set, as most learners of Japanese must, the frequency with which new characters and words are encountered has a significant impact on acquisition. A skill specific to Japanese learning is the ability to guess kanji words from context. Interestingly, most kanji have a component that carries semantic meaning, and similarly to guessing a meaning of a new word in English by understanding a Latin or Greek root, kanji meaning can be guessed through a combination of orthographic and contextual clues. However, Mori and Nagy (1999) found that only half of the Japanese learners they tested could use an effective combination of morphological and contextual information, and that most readers relied on one more than the other. The best way to learn kanji is through exposure to the characters. In a study of adult learners of Chinese, which uses the same

characters, the learners were able to recognize and pronounce high-frequency words with significantly more accuracy than low-frequency words (Sergent & Everson, 1992).

Japanese learners are exposed to kanji characters with less frequency than Chinese learners because many classrooms and textbooks limit the amount of kanji that Japanese students are exposed to, choosing to use kana instead. Eveson and Kuriya (1998) found that, when faced with kanji in an authentic text, many readers showed signs of dismay. The readers seemed demotivated by the difficulties they encountered interpreting kanji words, and some seemed to avoid the kanji altogether. As Japanese words can be written in either kanji or kana, many texts aimed at learners choose to use kana in order to be more approachable.

Furigana, phonetic annotations of kanji words, may be the answer to this problem. Taylor and Taylor (1983) found that words that appear in between lines of text are still registered by the reader, even if their attention is not on it, and Kirwan (2003) argues that this indicates a possibility that even if a reader's full attention was on the furigana, they would still subconsciously register the kanji. Kirwan (2003) conducted a study on a group of beginning Japanese students. By switching from a kana-only textbook to one that included kanji with furigana, even though the students were not directed to study the kanji and were expected to read using only the furigana, the students were able to recall 7% of the kanji they were exposed to by the end of the term. While the number is small, they were still able to learn kanji without any explicit teaching. This finding means that text for learners that is comprehensible can teach kanji as well as vocabulary implicitly, without the learners having to consciously study kanji characters-- a task that is generally considered one of the most difficult, tedious, and de-motivating aspects of learning Japanese.

When teaching a second language, it is important to consider that in order to

comprehend a piece of text, the reader uses a combination of lexical processing, syntactic processing, and visual processing. Each kind of processing presents unique challenges for L2 readers, and understanding these differences is key to understanding how one can make a text approachable for L2 readers. If the most important component of reading comprehension is adequate sight vocabulary coverage, then a text simplification system such as Dokusha should prioritize lexical simplification. It seems reasonable to aim for 98% known words in a piece of text aimed at L2 readers, and to present these words in their orthographically typical form, with furigana to aid in pronunciation. Complex sentences with multiple subjects or garden-path sentences which require more cognitive processing are prime candidates for syntactic simplification. The more fluently the learner can read the text, the more they are able to acquire new vocabulary, so any step taken to increase the readability of the text will enhance learning for the L2 reader.

Special Considerations for Teaching L2 Literacy

Second language readers may not be able to rely on the same beginning reading strategies they used when acquiring their L1 in childhood. While it may be important to understand how a word represented in text should be pronounced for acquisition or comprehension, L2 readers don't have a large enough vocabulary to benefit from ‘sounding out’ a word in the way L1 readers do. Even the most common L1 reading strategies, such as looking for semantic clues in the context of the sentence and paying attention to discourse markers, may be very difficult for L2 readers unless they understand the vast majority portion of a text.

Phonetic Processing in L2 Reading

Beginning L1 readers have a large auditory receptive vocabulary from the 4-5 years they spend learning their native language orally before learning to read. By the age of six, a native English speaker will have somewhere between 5,000 and 7,000 words, with some estimates as high as 8,000 (Carey, 1978; Farrell, 2009). To capitalize on this wealth of knowledge, L1 literacy teachers place emphasis on the relationship between phonemes (sounds) and graphemes (spelling). Learners are encouraged to 'sound out' the word, with the idea that they will understand that they have said the word correctly when the sounds they make trigger recognition of a word already in their vocabulary (Stanovich, 1980).

Second language learners, by contrast, lack the lexical bank of the native speaker. Thus, asking them to sound out the word to trigger recognition of the word is often ineffective. If a student does not know a word orally, then they will be unable to recognize that word when it is sounded-out (Farrell, 2009; Grabe & Stoller, 2013). However, if the phonetic information of an unknown word is unclear, the chance that the reader will be able to use that word when writing or speaking may decrease (Mori, 1998; Senoo & Yonemoto 2014). This is especially true for Japanese learners whose L1 is a phonetic language such as English.

Sounding out a word is not without merit; but its benefits may be limited to more advanced learners of a language. Kondo-Brown (2006) found that readers who encountered unknown kanji in an authentic text had a better chance of inferring the meaning of the word if they were able to guess the pronunciation of the word, at least in part. The reasons why a native English speaker's knowledge of the pronunciation of a Sino-Japanese character influences their ability to guess a word's meaning are unclear. Kondo-Brown hypothesized

that having the ability to guess a character's pronunciation makes the reader feel able to work out the meaning, thus encouraging them to try harder. If a reader who relies on phonetic information to process words is unable to guess at the phonetic information encoded in a character, they may be unlikely to use contextual clues to theorize about the character's semantic information.

The Language Threshold Hypothesis

Despite the differences in L1 and L2 learning, L1 reading skills do play an important role in some aspects of L2 reading. A learner reading in their second language will process the text using both (or all) of their languages to some degree. However, there has historically been little attention paid in the literature to exactly how the cross-linguistic interaction from a reader's L1 affects L2 processing of text (Muljani et al., 1998).

The native language of a reader will influence everything from their rate of syntactic processing to their motivation for reading (Koda, 2007, 2008; Scott & de la Fuente, 2008). Although most studies concerning the effect of L1 processing on L2 reading have focused on the degree to which L1 reading skills carry over, there is a possibility that the languages interact during processing, and that students from different L1s may process the same L2 in different ways (Muljani et al., 1998).

When reading in their first language, readers are focused on discourse markers and may have developed reading strategies to extract the appropriate information from text. Before these reading strategies can be applied to their second language, they need to have accumulated enough language that they are not struggling to decode the vocabulary and grammatical structures of the text. As discussed previously, researchers have converged on

98% coverage of the running words of a text as a probabilistic threshold beyond which a reader should have little difficulty with the decoding process, enabling them to utilize cognitive resources to interpret the text using the reading processes they have practiced in their L1. When a reader is past the linguistic threshold of a given piece of text, that text is said to be comprehensible input for the learner, and at the ideal level to aid in language acquisition.

Older learners begin to read in their L2 when they are fluent readers in their L1, having already learned literacy skills in their native language. These skills do not automatically transfer to L2 reading. Clarke's (1980) Language Threshold Hypothesis makes the argument that accessing L1 reading skills requires that the student have a sufficient amount of L2 knowledge. Once this threshold has been passed in the L2, the L1 reading strategies can be used to interpret the text. This means that while second language learners are often proficient first language readers, they may not be able to use the same reading strategies in their new language until their second language skills have reached a critical mass for the target text. While L2 knowledge may take a variety of forms, for the purposes of most research it is defined as the lexical and syntactic knowledge of the learner.

The Language Threshold Hypothesis, built from a combination of Clarke's (1980) study and Cummins' (1979) work with bilingual children, states that below a certain L2 proficiency there is little positive benefit from L1 reading abilities on L2 reading, and that above this threshold there is a positive correlation between L1 reading abilities and L2 reading abilities (Grabe & Stoller, 2013). In Clarke's study on Spanish speakers learning beginning English, the readers were divided into groups based on Spanish reading ability. While the difference between the two groups was clear on a Spanish reading comprehension

test, when asked to perform an English reading comprehension task, the 'good' readers had only a small advantage over the 'poor' readers (Clarke, 1980). These results resulted in a theory that low L2 skills "short circuit" a normally proficient reader's methods, leaving them unable to use the strategies they would employ to understand a text written in their L1. Lee and Schallert's (1997) findings from their study on the L1 and L2 reading skills of Korean students support the Language Threshold Hypothesis. They found that students in the high L2 proficiency groups showed a positive correlation between L1 and L2 reading ability, where the students in the lowest 30 percent of the students showed little influence of L1 reading ability on L2 reading comprehension.

An important caveat of the Language Threshold Hypothesis is that it does not indicate that there is a single threshold for all readers and all texts. Rather, the hypothesis suggests that, for a text to be at the optimal level for a reader to be able to access the literacy skills developed in their L1, the reader must know almost all the words and grammatical structures encountered. As the cognitive resources of L2 readers are diverted towards decoding the language of the text, there are few cognitive resources left for the L1 reading skills to be employed (Grabe & Stoller, 2013). The reading skills that are negatively affected by a low L2 proficiency include the ability to infer the meaning of unknown words from context, a key component of vocabulary acquisition through reading, as well as discriminating between the main idea and extraneous information in the text (Laufer & Ravenhorst-Kalovski, 2010). By crossing the linguistic threshold, or by the level of the text being lowered enough for the learner to rise above it, the cognitive resources that may have been used to decipher language structure and vocabulary are free to aid in the fluent comprehension of the text, which leads to greater learning gains (Pulido & Hambrick, 2008).

The Comprehensible Input Hypothesis

Krashen's (1985) Comprehensible Input Hypothesis states that the only thing a learner needs to acquire a language is comprehensible input at an appropriate level. Input is defined as text to read or speech to listen to. Krashen describes language as a linear series of steps each learner goes through. If a learner is at step 255, they would need input at step 256 to acquire the target language at an optimal speed. Using this logic, comprehensible input is defined as i + 1, where i is representative of the current level of the learner. The Comprehensible Input Hypothesis claims that the only thing necessary to move a learner from i to i + 1 is input at an i + 1 level. Krashen suggests constructing class content around comprehension based activities even from the very beginning. This means that instead of teaching grammar directly, Krashen advocates that teachers should produce model language for the students to listen to, and direct the students to level-appropriate reading material.

At advanced levels, Krashen (2003) suggests utilizing "sheltered subject matter teaching", which the author defines as learning a different academic subject in the target language with other intermediate language learners of similar proficiency. Hauptman, Wesche, and Ready (1988) showed the efficacy of this approach when they studied the effects of teaching second-semester psychology in French to a group of French language learners. The teachers were psychology professors, as opposed to language teachers. At the end of the class, the students were at the same level of French proficiency as the students that spent the semester in a traditional French language classroom, and their knowledge of psychology matched those who took the same psychology course in their native language.

While some researchers disagree with the idea that language production is not necessary for language acquisition (Izumi, 2003), the literature mostly agrees that reading is

an important component in any language curriculum, with the caveat that learning occurs optimally at a particular relative difficulty level of text (Nation, 2013).

When taken together, both the Language Threshold Hypothesis and the Comprehensible Input Hypothesis indicate that there is an ideal level for text to be at in order for optimal language acquisition through reading to occur. Text should be at a low enough level to be read fluently. The disruptions caused by working through complex sentences and new vocabulary shouldn't overtax working memory, causing the overall meaning of the sentence to be lost. However, the ideal reading material for the learner should not be so easy that there are no new words, kanji, or grammar structures present. If learners are not continuously exposed to language that is just beyond their current vocabulary, their language will stagnate and no learning will occur.

The Potential of Technology for Generating Level-Appropriate L2 Text

Natural language processing is the subfield of computer science that deals with the understanding and generation of human language by a computer. Recent advancements in the field of natural language processing unlock the possibility of automating a shift in the difficulty level of a piece of text, as opposed to having an appropriately-leveled text rewritten from an original source text by a human author (Biran, Brody, & Elhadad, 2011; Mino & Tanaka, 2011; Nunes, Kawase, Siehndel, Casanova, Dietze, 2013; Paetzold & Special, 2013, 2015; Shardlow, 2014b; Tsuchiya & Sato, 2003). Accomplishing this requires a combination of computational processes that work in tandem to reduce a text's complexity while maintaining its coherence. One of the most visible applications of natural language processing is machine translation between human languages, as used by products such as

Google Translate (Google, n.d.). The basic processing behind machine translation between languages could also be used to translate from complicated language to simple language. Within the field of natural language processing there are further subfields of machine translation, information extraction, text simplification, and summarization (Jurafsky & Martin, 2008). The techniques from these subfields can be used to build software that can rewrite a piece of text to make it more accessible: simpler vocabulary, less complex sentences, and shorter text overall.

The problem of reducing the vocabulary load of a piece of text can be addressed through the use of lexical databases. Lexical databases connect words in one language to those in another or describe relationships between words within a language, such as the Princeton WordNet project. Such databases could be used to replace difficult words with more commonly taught alternatives. The Princeton WordNet is a lexical database that started by documenting the relationships between English words, and has since spread to many other languages (Miller, 1995). It contains an entry for most words in the English language, and each word may have multiple "senses", each corresponding to a different meaning of the word. Multiple words may share a sense, and this shared sense is represented in WordNet as a "synset". While the words in a synset may not be perfectly interchangeable -- they may have different connotations, or be used in more or less formal situations -- they still share a core meaning. For example, the word 'exercise' has multiple senses. One sense is in a synset that includes words such as 'practice' and 'drill'. Another sense is in a synset that includes the words 'physical exertion' and 'workout'. The Japanese version of WordNet was developed in 2006 as an extension of the Princeton WordNet; Japanese words were added to the original English synsets (Isahara, Bond, Uchimoto, & Utiyama, 2008). This means that,

as long as the correct sense of a word can be ascertained from context by a computer, WordNet can be used to automatically locate potential synonyms.

Reducing the grammatical complexity of sentences involves removing extraneous phrases, or converting long sentences into multiple smaller, simpler sentences after they have been parsed by a syntactic dependency parsing system. In a dependency parse, each word or phrase in a sentence is connected to the parts of the sentence it modifies, building up a tree of grammatical relationships. Among other things, this can be used to identify the subject or object of a given predicate in a sentence (Jurafsky & Martin, 2008). Because of the structure of Japanese, it is particularly unclear what constitutes a word, and instead of words, the dependency trees are often built with a grammatical concept known as a "bunsetsu". A bunsetsu is made up of an independent word -- such as a noun, verb, or adjective -- and accompanying words -- such as postpositions, or auxiliary verbs. Japanese dependencies are simpler than dependencies in English, as each bunsetsu depends on a bunsetsu somewhere later in the sentence, with the final bunsetsu being the head of the dependency tree (Kurohashi & Nagao, 1994).

太郎は	りんごを	食べた	花子と	会う
Taro TOPIC	**apple OBJECT**	**ate**	**with Hanko**	**meet**

Figure 1: A sentence segmented into bunsetsu

English has more complicated dependency relationships, as the predicate, which is most often the top of a dependency tree, is usually in the middle of an English sentence. With the dependencies built up into a known hierarchy, a system could remove extraneous phrases from a sentence while keeping the main idea, which would be found further up the dependency tree, thus reducing the complexity of the sentence while preserving its essential

meaning. Dependency trees have been used to break long sentences apart into two smaller, more manageable sentences (Hayashi, 1992). The basic idea behind this technique is to take a full branch of a dependency tree and convert it into its own sentence.

The length of the text can be adjusted using automated summarization strategies. In the context of natural language processing, text summarization is the automatic distillation of the most important information in a given text. There are two major kinds of text summarization being researched today: extract and abstract summarization (Jurafsky & Martin, 2008). The simplest form of summarization is an extract, a summary produced by pulling the most important information from a document word for word. By contrast, an abstract is created by generating new language to summarize the contents of a text. The extractive summary is much easier computationally than the abstract summarization, and consists of three major tasks: selecting the relevant content, ordering the extracted phrases, and rectifying coherence problems in the patchworked summary (Jurafsky & Martin, 2008).

In addition to reducing the level of the language, other pedagogical strategies can be applied automatically by software; such as annotating the text with definitions of relevant words, or, in the case of a language like Japanese, altering the way a word is represented to make it easier to recognize or acquire. These automated changes to the appearance of the text could scaffold the text for a reader, enabling them to read more complicated language than they could otherwise read unassisted.

Providing short definitions or multimedia representations of difficult words in the text, either in the reader's native language or the target language, is called glossing. Glossing difficult words in the text requires the correct sense of the words to be extracted from the context of the sentences they appear in. Once the correct sense of the word has been

disambiguated from the other senses, the correct definition, either in English, simple Japanese, or a multimedia gloss, can be presented in the margins or in a modal popover, or even embedded in the text itself (Shardlow, 2014b).

The way in which a word is visually represented in the text could make that word more recognizable to Japanese learners. Japanese is made up of two syllabaries (hiragana for native words, katakana for loan words) and one set of logographic characters from Chinese (the Sino-Japanese kanji). A combination of the two can be constructed using ruby text (furigana), which places the phonetic pronunciation in hiragana above the logographic kanji. This could be especially important for allowing readers to take on less familiar words, as without readily understanding the pronunciation of a word, it will be difficult for the reader to acquire it, even if they can correctly guess the meaning from context (Senoo & Yonemoto, 2014).

If a computer program could be constructed to reduce the level of a piece of text automatically, then it would eliminate the problem of finding a large enough variety of appropriately leveled text for a Japanese extensive reading program. Additionally, independent learners would be able to self-convert text found online to the ideal level for them to read fluently for meaning. It would even be possible for the same text to be scaled to multiple levels, enabling learners to work their way into more complicated language with a solid understanding of the content of the text obtained through initial readings at lower difficulty levels, or for students at different reading levels in a group class to work through the same text at the level most appropriate to their current L2 learning needs.

To demonstrate that techniques such as dependency parsing and automatic text summarization could be used to generate language learner targeted literature, the design,

implementation, and assessment of the software project Dokusha will be examined. Dokusha is a software program that takes a Japanese text and reduces its level to one appropriate for a student at one of the 5 levels of proficiency designated by the Japanese Language Proficiency Test (JLPT) using the techniques outlined above (The Japan Foundation, 2018). As second language acquisition research turns more towards fostering autonomy and literacy and natural language processing resources such as WordNet (Miller, 1995) are more abundant and complete than ever before, the goal of Dokusha is to demonstrate that this confluence of advancements in the field of natural language processing can be used to scale the grammatical and lexical complexity of a piece of text to a level appropriate for second language learners at various stages of language proficiency.

The goal of Dokusha is to generate text where the percentage of known words hovers as close as possible to the 98% goal found by reading researchers. Where many text simplification systems are written with the goal of simplifying the text to the simplest possible form that retains the same meaning, Dokusha targets this concept of an optimal level more specifically, leaving enough new words to grow the receptive vocabulary of the reader.

Features, Functionality, and Goals of Dokusha

The need for different teaching strategies for L2 versus L1 readers highlights the need for literature geared specifically toward foreign language learners, and guides the design of features for Dokusha. 'Dokusha' means 'reader' in Japanese, and the aim of Dokusha is to automatically generate text at one of the five JLPT levels of Japanese from authentic Japanese input text, such as news articles. Dokusha's automatic leveling system will be implemented primarily using automated text simplification. Text simplification is a sub-field

of natural language processing that modifies natural language, reducing linguistic complexity in order to increase readability. The techniques used in automated text simplification also draw on related fields discussed above, such as machine translation, text summarization, and paraphrase generation.

As discussed in the previous section, the two main challenges to be mastered when reading a second language are vocabulary and grammar. Text simplification can be broken into two main areas of research accordingly: lexical simplification and semantic simplification. Lexical simplification aims to simplify the vocabulary used in the text. Semantic simplification untangles complex grammar to make the text easier to read. To reduce a text's lexical and semantic complexity, Dokusha must first be able to understand the level of the vocabulary and grammar of the original text.

Assessing Original Text Level

The first task of text simplification processes is to identify what about the text must be simplified. If the goal is to provide a level-appropriate text for a reader, then it is not useful to simplify everything down to the lowest possible level, as some simplification systems aim to do. Instead, the level of the reader must be established and the correct level of text produced. This means not simplifying sentences if they are already comprehensible to the target audience. Sentences that are simplified also may not need to be converted to the simplest form-- only grammar and vocabulary the student has not yet been exposed to, or would be unable to guess from context, should be addressed. While text presented at too high a level demotivates and disempowers the reader, oversimplification of the text under-

challenges the reader and will not allow the reader to develop their reading comprehension skills.

Understanding and quantifying what precisely defines any given ‘reading level’ is a complex and multidimensional problem. To simplify this task, Dokusha will use a system of established levels based on the JLPT (The Japan Foundation, 2018). The JLPT is a test administered internationally to measure and certify the level of Japanese proficiency of non-native speakers. There are five levels of the JLPT, with N5 establishing a basic understanding of simple sentences, and N1 being near-native level fluency.

The list of kanji, vocabulary and grammar tested at each level has not been officially published since the test switched from its previous 4 level system to the current 5 level system in 2010. Various groups have compiled lists based on the outdated lists from 2009, and from the content of the tests that have been administered since. Dokusha uses these lists to assign a level from 0-5 to every word encountered in a text, with 0 representing the words that don't exist on any JLPT study list and are therefore considered the most difficult, and level 5 being the easiest. The words that are flagged with a lower integer than the target level are more difficult than the target level for the text are flagged for substitution or removal.

Lexical Simplification

Lexical simplification is a multi-stage computational process that generally follows the following four steps: Complex words are identified and flagged for removal, lists of substitutes are generated, the substitution candidates that are inappropriate for the context are filtered out, and finally, a synonym is chosen and swapped into the sentence in place of the existing word (Shardlow, 2014b).

Complex word identification. In Dokusha, the first step, complex word identification, is done using the JLPT word lists as discussed above. This simplifies the initial step of the lexical simplification process. Any word that is neither a proper noun, nor contained in the JLPT vocabulary list at or below the target level will be considered a complex word, and will become a target word for further steps.

Synonym generation. The next step takes each target word and compiles a list of all possible substitutions. This is typically done in one of two ways. Either the word is looked up in a pre-made lexical database that ties words to other words that share a meaning, or synonym candidates are generated in this step from other sources such as dictionaries or corpora. The best approach may be to use a combination of approaches in order to generate the most comprehensive list of possible substitutions, which is what Dokusha does.

The most frequently used lexical database for lexical simplification is Princeton's WordNet (Miller, 1995; Paetzold & Specia, 2017). Each word in WordNet is tied to one or more synsets -- groups of words with the same sense in common. Looking at the other words in a synset will give synonyms for a given target word. In addition to synonyms, some researchers have begun using hypernyms and hyponyms (Biran et al., 2011). A hypernym is a word that can be described using an is-a relationship to the target word. For example, fruit is a hypernym of apple, as ‘an apple is a fruit’ is a true statement. Conversely, it follows that apple is a hyponym of fruit. While the hypernym will not provide as much specificity of information as the hyponym, it is likely to be easier to understand.

Kajiwara, Matsumoto, and Yamato (2013) presented an alternative strategy for

generating synonyms based on dictionary entries from a Japanese dictionary. The target word is looked up in the dictionary, and the resulting definition is parsed and tagged with parts of speech. The words that match the part of speech of the target word are considered to be candidates for synonyms. Then the words are ranked by their semantic distance to the target word, calculated using WordNet (Ma, Fellbaum, & Cook, 2010). One of the advantages of using dictionary definitions is that definitions generally use simpler language, so any word that appears in the dictionary definition is likely to be simpler than the search term (Mino & Tanaka, 2011).

Glavaš and Štajner (2015) proposed an automated way of generating synonyms that does not rely on WordNet or dictionaries using the cosine similarity between word vectors. Each word is turned into a vector which numerically quantifies the word, allowing it to be compared to other words in the language. There are multiple ways to convert a word into a vector, each of which allows for different kinds of comparisons between words. Two such vectorizations, co-occurrence vectors and word-embedding vectors, allow the concept of "closeness" in meaning and therefore of suitability for substitution to be quantified by a mathematical distance. To accomplish this, the cosine of the angle between vectors is calculated, leaving a number between -1 and 1, with numbers closer to 1 indicating a higher degree of similarity. There are two main types of vector, both pre-calculated on a corpus.

A co-occurrence vector has as many dimensions as there are unique words in the language, where each dimension is a number which represents the number of times that word appears near the target word. "Near" is defined as how many words away an occurrence of a word is from the target word, and the level of "nearness" examined by the co-occurrence vector-- that is, how many words around the target word the co-occurrence vector examines--

is defined by the window size of the co-occurrence vector. For example, a co-occurrence vector with a window size of 10 will contain counts for occurrences of words that appear in the five words before or after the target word every time the target word appears in the corpus. To reduce noise, often some threshold is defined below which occurrences of a word will not be counted. For example, some authors have chosen to disregard all words that only appear twice near the target word, setting that variable to zero (Biran et al., 2011). It is important to note that when the cosine similarity between two co-occurrence vectors is calculated, the words that appear near one word but not the other are eliminated, and that the furthest away two words can be is 0.

Word-embedding vectors are mappings of unique words into a vector space of predetermined dimensions, often using shallow neural networks. A common approach is word2vec, pioneered by Mikolov, Chen, Corrado, and Dean (2013). Word2Vec is a shallow neural network that operates on one of two modes: CBOW or skip-gram. In the continuous bag of words (CBOW) model, the vector represents a series of weights that when applied to a projection space of the vocabulary give the probability of a word given the surrounding words (unordered). The skip-gram model uses the target word to calculate the probability of the surrounding words, with the order specified.

Glavaš and Štajner used a 200-dimensional CBOW vector representation of all content words. The synonym list proposed by their system consists of the 10 words with the smallest cosine distance to the target word, excluding morphological derivations of the target word, such as ‘quiet’ and ‘quietly’. In a comparative study of different approaches to substitution generation, this unsupervised method of using raw text proved to be very competitive with more resource-heavy approaches (Paetzold & Specia, 2017).

Each of the above approaches has different strengths and will find different synonym candidates. Therefore, the decision must be made: Is it most important to have more candidates, increasing the chance that a good match is among them, but making the process of selecting that match more difficult due to the larger pool, and requiring a third step to select the most appropriate match from among the candidates? Or should an approach be favored which keeps the pool smaller, may not find the most ideal candidate as often, but requires less whittling down of potential replacement words, thereby shortening or even potentially eliminating the selection step? Paetzold and Specia (2017) showed that a combination of approaches could greatly increase the probability that an appropriate synonym for the target word could be found in the list of synonyms generated. This is intuitive because an additive approach will generate more synonyms overall. As Dokusha has a disambiguating step meant to eliminate semantically dissimilar words, the 'whittling down' is left to that process. This will allow the overall goal of this second step to be to create a large list of possible candidates, giving the highest probability that an accurate substitute word will be found.

Filtering the candidate list. The next step in the lexical simplification process takes the generated list of substitution candidates and determines which subset of these words will fit the context of the sentence. This is the step that will ensure that the sentence will retain the same meaning as the original sentence when the lexical simplification process is finished, and will result in a smaller, culled version of the candidate list generated by the previous step. The first published lexical simplification systems skipped this step, but Shardlow (2014a) studied lexical simplification errors on 115 sentences and found that nearly 30% of

the errors could be attributed to the chosen synonym altering the meaning of the original sentence, highlighting the need for this step to be taken if the accuracy of automated text summarization is to be increased.

In some sentences an original word is too complex for the target reading level, but cannot be replaced. For example, in the sentence "At Christmas I eat pandoro, a traditional Italian cake", the word 'pandoro' is likely to be flagged for simplification. However, the most appropriate synonym to replace it with would be 'cake'. This would be fine in a sentence such as "I ate too much pandoro", because the simplified sentence "I ate too much cake" would carry the same meaning. But in the first sentence, the word 'pandoro' is being defined within the sentence, so it should remain. To solve this problem Biran et al. (2011) took two steps. The first was to make sure the synonym chosen for simplification was not already in the sentence. The second method was to create co-occurrence vectors both for the target word in the context of the sentence, and for the target word in a corpus. The cosine similarity of the two vectors was calculated and if the similarity was higher than 0.1 (a threshold found through experimentation), the sentence was left un-simplified. This was done because if the cosine similarity between the two vectors is high, then the target word is being used in the ideal context for that word, and should not be substituted.

If the target word is swapped out for one of the substitutions, one of the following two problems may arise: Either the sentence may lose the meaning carried by the original, or the sentence may become ungrammatical. To lessen the possibility of grammatical errors, the substitution candidate list is often filtered to only include words that share the same part of speech tag as the target word (Paetzold & Specia, 2017). This may not help preserve the meaning of sentences in many cases. Paetzold and Specia (2013) found that only 56.5% of

the sentences they simplified retained the same meaning after using a similar approach.

To ensure that the new sentence will have an equivalent meaning to the original, only the substitution candidates that preserve the meaning of the sentence should be selected. One approach to this selection task is to use word sense disambiguation, which uses classification to choose the correct WordNet sense of the target word in the context of the sentence. As WordNet has separate synsets for each sense, if the correct sense of the word in context is chosen, then the words that appear in the same synset as the target word in WordNet, or are closest to the target word sense in WordNet similarity, can be chosen (Nunes et al., 2013). This approach is limited by the reliance it builds on sense databases such as WordNet, and limits the usefulness to single word simplification as determining the sense of a phrase is much more difficult. Additionally, understanding which WordNet sense is being used by the target word in the original sentence is often a classification task which relies on training data that has been manually disambiguated.

Other approaches besides word sense disambiguation for filtering the list of substitution candidates use word vectors. Biran, Brody, and Elhadad (2011) created a common context vector between substitution candidates and the target word vector. The common context vector is a vector that uses the minimum value for each feature between the candidate word and target word. This means that words that are commonly found within the 10 word window of the candidate word, but not the target word, and vice versa are disregarded. The cosine similarity between this common context vector and the context vector of the sentence is calculated and the highest candidate selected. Paetzold and Specia (2015) proposed a similar vector based model, but used a word-embedding vector instead of a co-occurrence vector. Instead of a threshold, they eliminated the half of candidates that

scored the lowest. In a direct comparison between the two approaches, Paetzold and Specia's approach was found to have the higher precision and recall than Biran et al's (2011) (Paetzold & Specia 2017).

Other techniques for using context to disambiguate senses use n-gram models directly or indirectly. N-gram models train on a corpus and work by observing the pattern of words that are found together. A 3-gram model, for example, would look at every group of three words that appear consecutively. The sentence "In the Edo period, life was quite monotonous" consists of five 3-grams: "In the Edo", "the Edo period", "Edo period life", "period life was", "life was quite", and "was quite monotonous".

In an example of indirect n-gram model use for word sense disambiguation, De Belder and Moens (2010) used a latent variable language model to implicitly choose synonyms that share the same sense, eliminating the reliance on WordNet for choosing synonyms. The Latent Words Language Model uses a machine learning approach that trains a Bayesian network on an unannotated corpus, creating a hidden layer build on observed 3-grams. The latent variables are the synonyms and related words. The candidates that also appear in the latent words model are chosen, and the rest eliminated. This approach comes at a high computation cost, despite having high accuracy.

Another approach built on the n-gram model was used to analyze the similarity of the meaning of a target word in context with a candidate word. Baeza-Yates, Rello, and Dembowski (2015) used the 5-gram consisting of the target word and the two words to the right and left of it. They swapped the candidate words in for the target word in the 5-gram and counted how many times the new 5-gram appeared in a corpus. The synonyms that

appeared in the 5-gram the most often were kept and candidates were eliminated if they scored below a certain threshold.

Simplicity ranking. The last step in lexical substitution is to take the final list of substitution candidates and choose one to substitute for the target word. This step is where the needs of the reader are most strongly taken into consideration, and many systems aim to choose the simplest of the remaining candidates.

The aim of Dokusha is to help non-native readers learn how to read better, not just more easily. Choosing the simplest word of the candidates for substitution might cause a reader's vocabulary and reading ability to stagnate at the lowest level the system is able to generate; thus, rather than choosing the simplest word of the candidates, the most level-appropriate candidate is chosen. This is done primarily with help from the JLPT word lists. If the reading is leveled to an N4 appropriate level, candidate substitution words that appear in the N4 word list will be given higher priority than words on the N5 list, increasing the reader's exposure to level-appropriate words. As these bands of words are fairly large, many ties between candidates with equal priority are expected with this ranking system. In the case where two candidate words appear in the same JLPT vocabulary list, Dokusha will default to selecting the candidate word judged by the previous step to be the most suitable for the given context.

Syntactic Simplification

Syntactic simplification is a technique for reducing the grammatical complexity of a text. This is done in two steps: First, the structure of the sentence is analyzed and it is

decided whether or not the sentence is a candidate for simplification. Second, the sentence undergoes a transformation, as it is rewritten according to a set of level-appropriate rules (Shardlow, 2014b).

Analyzing syntactic structure. The first step of syntactic simplification is to analyze the structure of the sentence. This is done by a parsing module that builds a parse tree from the sentence.

This can be done at high or low levels of the sentence. At low levels, each word is tagged with a part of speech and a tree that outlines the relationships between words in the sentence is drawn. At high levels, words are grouped together into larger chunks of the sentence. For Japanese, the lowest level (highest 'granularity') one may parse is that of morphemes, where each indivisible chunk of meaning is drawn in relation to the other morphemes in the sentence. The higher level groups together morphemes into 'bunsetsu', and the relationship each bunsetsu has to the others is represented in the parse tree. Syntactic simplification can use text that has been parsed at any level of granularity (Shardlow, 2014b), but the higher levels are easier to work with, so Dokusha parses sentences into bunsetsu, rather than into morphemes.

To do this, Dokusha uses a Japanese dependency parser made by researchers at Kyoto University called Kurohashi-Nagao-Parser, or KNP (Kurohashi, 1998). KNP takes a Japanese sentence and interprets its structure, also determining the sentence's case markers, defined by 'helper words' called particles that serve to isolate important pieces of the sentence, and performing coreference resolution, a process which resolves ambiguities that arise as a result of anaphora (such as pronouns and antecedents in English).

To do this, KNP uses the morphological analysis system JUMAN, also created at Kyoto University. JUMAN can take a sentence and split it into morphemes, each one tagged with its pronunciation and part of speech. JUMAN provides additional information about the meaning of each morpheme, such as a category or domain. For example, the word for "foreign country" is assigned the category "place" and domain "politics". KNP takes the output of JUMAN and analyzes the sentence construction to provide additional information about the sentence. KNP can group morphemes into bunsetsu and tag them with syntactic features. This is helpful for locating the predicates in the sentence, as well as determining their dependencies. For example, the sentence "Before the Edo period, people's everyday lives were very monotonous and every day was a repeat of the same thing", contains three predicates in Japanese: "monotonous", "the same" and "was a repeat". KNP finds that the subject of "monotonous" is "lives", the time period for "monotonous" is "before the Edo period", and the modifier is "very".

This first step also decides whether or not a sentence should be simplified. Dokusha chooses to simplify sentences that contain a grammar point that is beyond the target JLPT level set for the text. The JLPT has a set of predetermined grammar that is tested at each level, which reduces the choice of whether or not to simplify into whether or not a grammar point higher than the desired JLPT level has been detected. Tsuchiya and Sato (2003) used a rule-based system based on the JLPT test specifications from before 2010 to assign a reading difficulty level to different parts of a given sentence. Dokusha uses a similar approach; each grammar point in the Nihongo So-matome textbook series (Sasaki & Matsumoto, 2010a-d) has a KNP rule associated with it, which is used to automatically detect the grammar point and its JLPT level during the parsing phase.

If a high level grammar point is found during the syntactic analysis of the sentence, the parse tree of the sentence moves on to the next step.

Sentence transformation. The second step of syntactic simplification is to transform the sentence from the original syntax, into the new syntax. This may take the form of replacing difficult grammar points with more basic ones, splitting long sentences into shorter ones, or rewriting passive sentences into active voice.

The grammatical manipulations to replace difficult grammar points are most often accomplished using handwritten rewrite rules, although some research has been conducted on automatically generating rewrite rules (Chandrasekar & Srinivas, 1997). Handwritten rules have the disadvantage of being work intensive, but the advantage of not relying on annotated corpora, and being more accurate (Shardlow, 2014b). For this reason, Dokusha makes use of handwritten rewrite rules.

For each grammar point in JLPT levels N1-N4, Dokusha has a KNP rule that will tag any bunsetsu that contains that grammar point, or is involved in that grammar point with an identifier and a level number. Sentences that contain higher level grammar points are flagged for simplification in the syntactic analysis step. The transformation step looks up the grammar identifier in a table of related grammar rules, where each grammar point has an approximate equivalent for each of the lower JLPT levels. The morphemes involved in the high level grammar are stripped out of the original sentence, and the new grammar rule is applied.

Dokusha splits sentences in two cases: the sentence is long and has at least two parallel bunsetsu, or the sentence is a garden-path sentence where a dependency for one of the predicates comes after the predicate in the sentence.

Klerke and Sogaard (2013) chose to split Dutch sentences where there were conjuncts as long as the split would result in at least one noun and one verb in each sentence. As Japanese subjects are optional in sentences, and often omitted, this would not be a practical approach for Japanese. Instead, long sentences that have two parallel bunsetsu that are more than three bunsetsu apart, are split. The sentence is ended after the first parallel bunsetsu, and the ending verb conjugated appropriately. The three bunsetsu in between ensure that the resulting sentences are not too short. If more than two bunsetsu are parallel in a single sentence, Dokusha works from the end of the sentence, splitting the last two parallel bunsetsu into separate sentences, then reanalyzing the new first sentence.

Automatic Scaffolding

The majority of research on the benefits of extensive reading have been conducted on English Language Learners, and not on learners of other languages, such as Japanese, where more a complicated orthography serves as an additional barrier to entry. Senoo and Yonemoto (2014) conducted a case-study on a single student learning Japanese. The student was provided with material to read for general comprehension, largely without the use of a dictionary, as well as material to read more intensively, looking up all the words the student didn't know. The student was able to make reasonable guesses as to the meaning of unknown words in the text when they were written in kanji because kanji are logograms that carry information about their meanings. The relationship between a kanji and its pronunciation,

however, is not so straightforward, and for rapid silent reading, there was no need for the participant to make a similar guess as to the pronunciation of the word. Without knowledge of pronunciation, the word cannot be learned incidentally to the degree that it could be used by the reader in a new context (Senoo & Yonemoto, 2014). This highlights an interesting distinction for Japanese Language Learners: vocabulary acquisition through extensive reading may require the pronunciation of unknown words to be provided so the reader need only guess at the meaning.

Glossing has received more attention recently as a way to assist learners in understanding a new word without guessing from context. This may be useful in the case that the reader does not understand enough of the context (below 98%) or to double check that the student is not misguessing (Laufer & Sim, 1985). The benefits of glossing over guessing from context is that glossing can be used to provide more accurate meanings for more abstract or specific words that would be difficult to guess from context, or that may appear only once. Compared to dictionary look-ups, glossing doesn't interrupt reading, so the focus of the reader can be on interpreting the meaning of the sentence as a whole (Nation, 2013). Selective glossing of words can also call attention to key vocabulary that readers may want to learn. Lee and Lee (2015) compared the efficacy of three different types of electronic glosses, made possible by ebook readers and other computer assisted language learning (CALL) programs. They found that tooltip glosses, where the definition appears in a small window near the target word, and frame-type glosses, where the definition appears in a separate section of the screen, were effective for reading comprehension and vocabulary acquisition.

In a hierarchical society that values ancient Japanese tradition, one cannot move up without intense training and learning the traditional conventions.

日本古来の伝統を重んじる縦社会の中で、しきたりを学びながら激しい稽古を積むことなしに、上に上がっていくことはできない。

Syntactic Analysis

日本古来の	伝統を	重んじる	縦社会の	中で、
しきたりを	学びながら	激しい	稽古を	**積むことなしに、**
上に	上がっていくことはできない。			

Syntactic simplification

日本古来の	伝統を	重んじる	縦社会の	中で、
しきたりを	学びながら	激しい	稽古を	**積まないで**
上に	上がっていくことはできない。			

Complex Word Identification

日本	古来	の	**伝統**	を	重んじる	縦社会	の	中	で	、
しきたり	を	学び	ながら	激しい	**稽古**	を	積まないで			
上	に	上がって		いく	こと	は	でき	ない	。	

Lexical substitution

日本	古来	の	**歴史**	を	重んじる	縦社会	の	中	で	、
習慣	を	学び	ながら	激しい	**練習**	を	積まないで			
上	に	上がって		いく	こと	は	でき	ない	。	

日本古来の歴史を重んじる縦社会の中で、習慣を学びながら激しい練習を積まないで上に上がっていくことはできない。

*In a hierarchical society that values ancient Japanese **history**, one cannot move up without intense **practice** and learning the traditional **customs**.*

Figure 2: Flow of Dokusha

Implementation of Dokusha

Dokusha's main function is to take authentic Japanese text and generate a simple version of the text at a target level. It does this sentence by sentence. First, authentic text is parsed and split into sentences, which are analyzed for length and shortened if necessary. Next, each sentence is run through Dokusha's syntactic simplifier, which reduces the grammatical complexity of the sentence. Following this stage, the grammatically-simplified text is checked for difficult words, synonyms are generated at or near the target level, and the lexical substitution step is performed. Finally, the simplified text, augmented with glosses and furigana, is returned. These stages will be explored in further detail in the following sections.

Splitting Parallel Clauses

First, long sentences with parallel clauses are split into shorter sentences. This is done through analysis of the dependency tree. If a branch of the dependency tree is marked as parallel to another branch, and the phrases are at least three bunsetsu away from each other, the sentence will be split. If a branch is marked as parallel to another, the two clauses can be separated into two sentences easily. The stipulation to have the phrases at least three away from each other prevents simple sentences with parallel constructions from being split. For example, it is not necessary to split a sentence such as "I ate apples, bananas, and carrots" into "I ate apples. I ate bananas. I ate carrots." It is, however, useful to split a sentence such as "Udon is a noodle made of white flour and while it has been made throughout the country since ancient times, not only in Kagawa, the Kagawan brand Sanuki is particularly well known, to the point that whenever a new brand is introduced, people call it the 'new Sanuki'

even if there is no relation to the original brand" into "Udon is a noodle made of white flour. It has been made throughout the country since ancient times. The Kagawan brand Sanuki is particularly well known. Whenever a new brand is introduced, people call it the 'new Sanuki' even if there is no relation to the original brand".

Syntactic Simplification

Once any overly long sentences have been split along parallel clauses, the shorter sentences are run through the syntactic simplifier, which looks through the dependency parse for any bunsetsu marked as containing a grammar point above the target level. In the case that a grammar point is spread across multiple bunsetsu, the bunsetsu are merged and dealt with as one. In figure 3 Below is an example of a high-level Japanese sentence's path through Dokusha's syntactic simplification system:

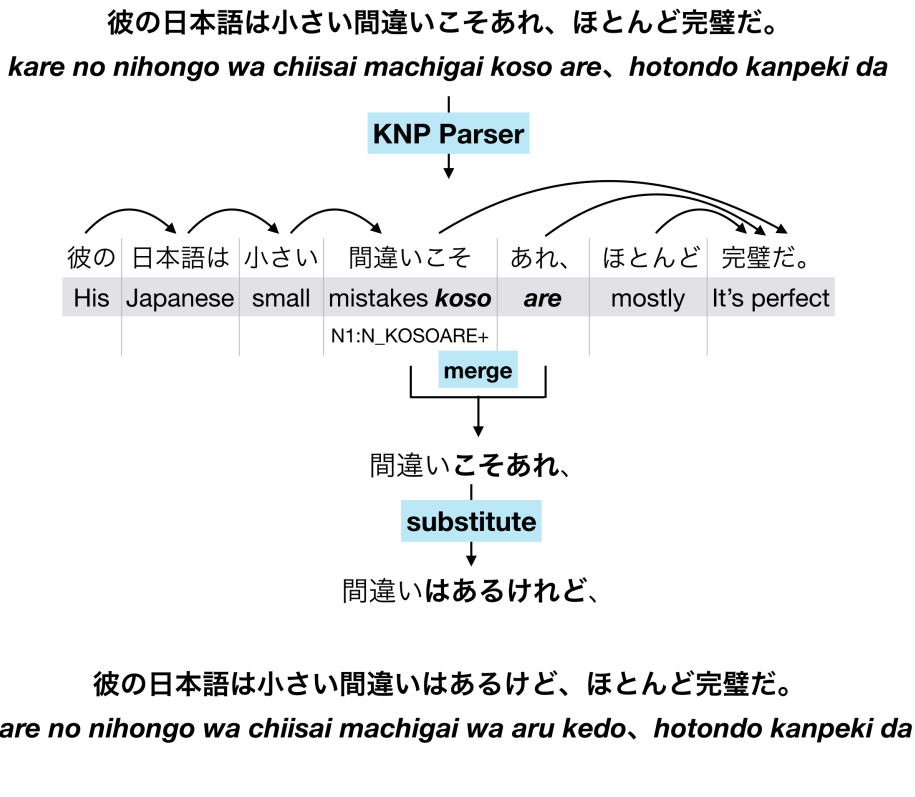

Figure 3: Syntactic simplification with merge

In this way, the original grammar point is removed from the bunsetsu and replaced with an easier equivalent. In the most simple case, the grammar point is constructed only of morphemes that are easily removed and replaced with new ones. Other grammar points may span multiple bunsetsu, the relevant morphemes separated by morphemes that must remain in the sentence, as in the sentence "The flavor and service are both bad at that restaurant".

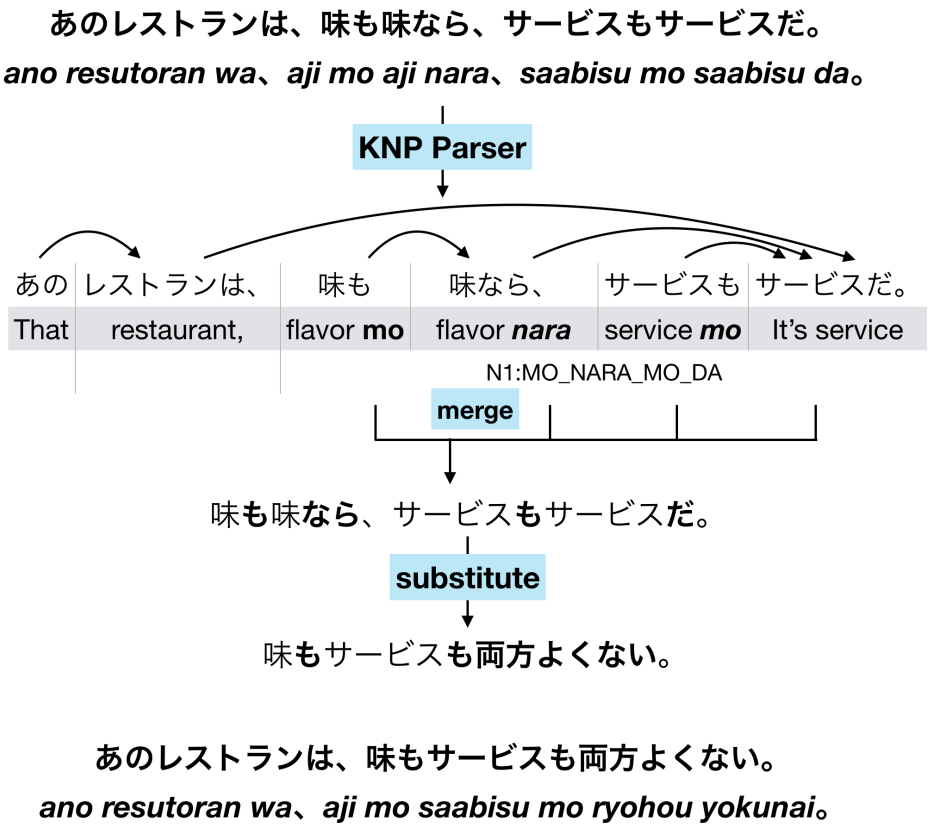

Figure 4: Syntactic simplification with multiple merged bunsetsu

Even more complicated substitutions require conjugation, as in the sentence "He paced in front of her house holding a bouquet."

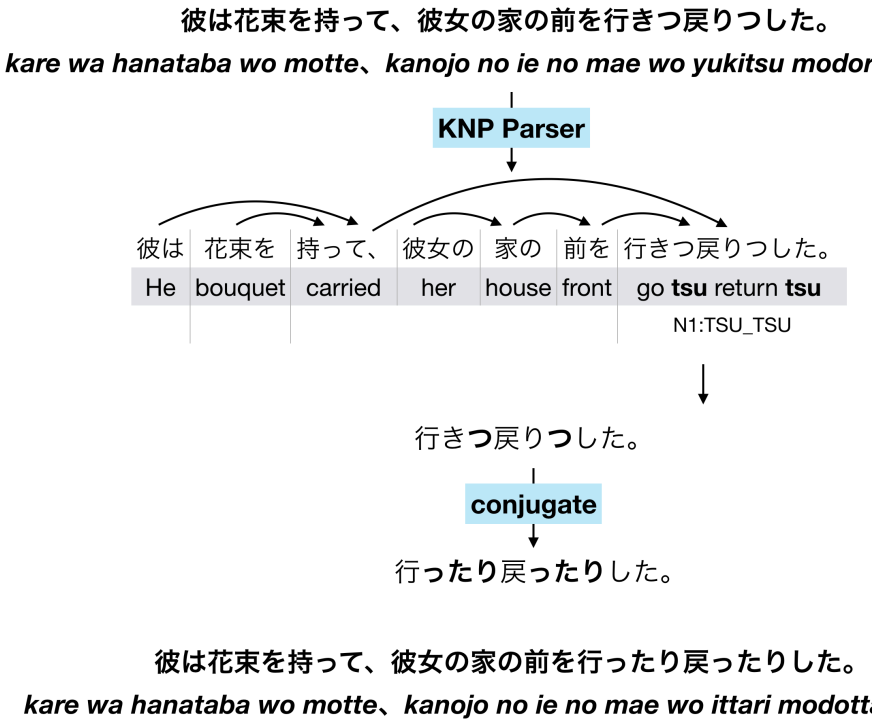

Figure 5: Syntactic simplification with conjugation

All of Dokusha's syntactic simplification is performed using handwritten rewrite rules. The textbook series Nihongo Sou Matome was chosen because the textbooks provide paraphrases rather than translations, meaning that for each high level grammar point, the textbook authors have written an equivalent phrase using lower level grammar points (Sasaki, & Matsumoto, 2010 a-d). During the development of Dokusha, the default KNP rule list was changed to include a new rule for each grammar point in the Nihongo Sou Matome textbook series. This new rule enables the grammar point to be detected within a sentence during the dependency parse, and communicates to Dokusha's syntactic simplification module which bunsetsu to merge, and which morphemes to remove from the original sentence. The grammar point is then mapped to a new grammar point depending on

the target level of the text. Any morphemes from the original sentence that must be conjugated are adjusted as necessary and the new morphemes are added to the sentence to complete the syntactic simplification process.

Lexical Substitution

Following the syntactic simplification step, the dependency parse of the simpler sentences goes through a process of lexical substitution. First, proper nouns are identified from the parsed sentence, because proper nouns generally cannot be substituted while retaining the original meaning of the text. Then any remaining content words (nouns, adjectives, adverbs, and verbs) that don't meet the target level requirements are flagged for substitution.

Synonym generation. The list of substitution candidates for a given target word is generated from several sources, as Paetzold and Specia (2017) showed that combining approaches for generating substitution candidates results in the greatest probability of a gold-standard candidate being found.

Vector model. The first method Dokusha uses to generate substitution candidates is the Glavaš and Štajner (2015) method. A 200-dimensional vector is computed using a continuous bag of words (CBOW) model over the Japanese wikipedia data. The fifteen most similar words to the target word are chosen and filtered by part of speech, with morphological derivations of the target word removed. These words are by definition the most contextually similar across the training corpus, but they are not always the closest in meaning.

WordNet. The WordNet synonym generator looks up the target word in WordNet and returns all the words that share the same synset as the target word. Because WordNet only links synsets that share the same part of speech, candidates generated this way are automatically the same part of speech as the target word. This approach improves the meaning accuracy of the substitution candidates. However, despite the words sharing a sense of the original word, they may not work in the same context.

WordNet plus hypernyms. This addition to the WordNet approach also returns the hypernyms of the target word, along with its synonyms as recognized by WordNet.

Word similarity metrics. Dokusha uses a number of word similarity metrics to rank the appropriateness of synonym candidates. Since some of the synonym candidates are generated based on similarity of context, and some are generated based on similarity of meaning, each candidate must be ranked by both metrics, so as not to be biased by the one that generated it.

Vector similarity. *V*ector similarity is measured using the cosine similarity of two word vectors. The vectors are constructed using word2vec (Mikolov, et al, 2013). Word2vec trains a shallow neural network to map words to vectors, which is then saved as a vector model. As discussed above, Dokusha uses two vector models, one trained on Japanese wikipedia article text and the other trained on the Balanced Corpus of Contemporary Written Japanese, with 200 dimensional vectors and a window size of 10.

The substitution candidates are each compared to the original word using cosine similarity, which produces a number between -1 and 1, where 1 is the closest two words can be. This number is subtracted from 1, so that 0 is the closest two words can be, and 2 is the

furthest. This is to match the next similarity metric, WordNet path similarity, which has 0 being the closest words can be.

WordNet similarity. WordNet similarity is a measurement of how similar two synsets are in meaning. The simplest metric is ***path distance***, which counts the number of steps along a path from one synset to another. For example, there are four nodes between the synset for 単調 (monotony) and the synset for 単純 (simpleness), as shown in figure 6 below.

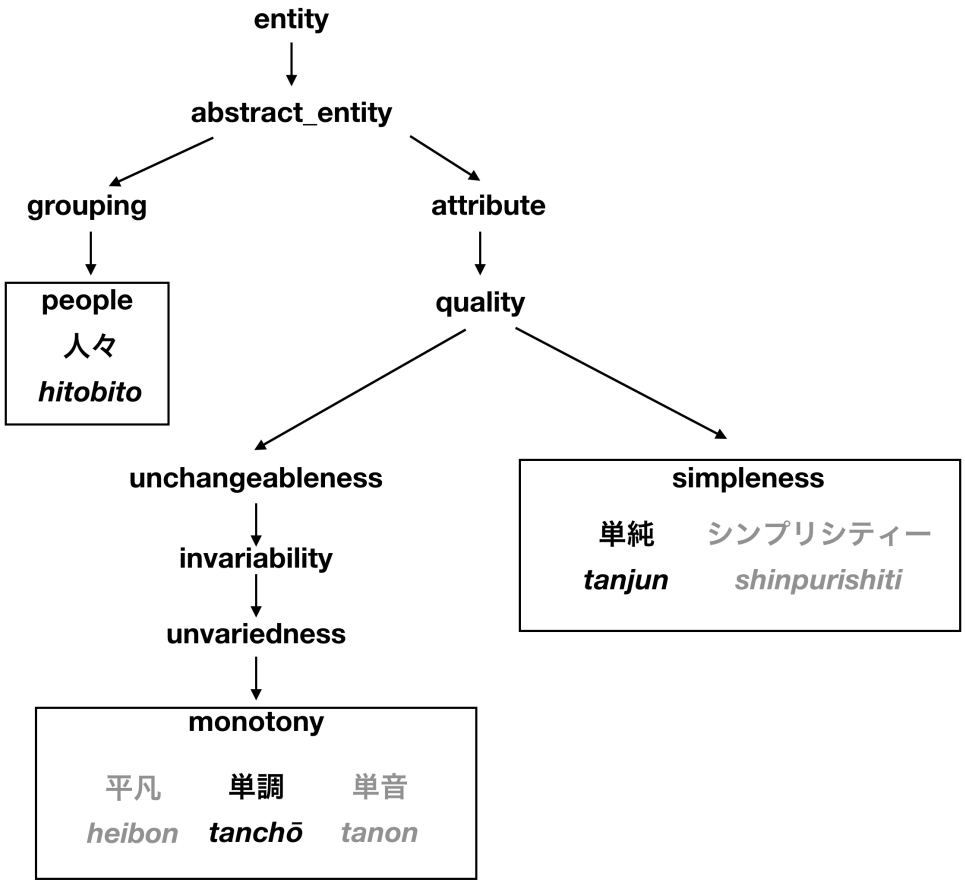

Figure 6: Partial WordNet visualization

Path-based similarity metrics look at the tree generated by hypernyms and hyponyms in WordNet. Following the hypernym path up from any given noun will inevitably lead to 'entity' as the overarching concept.

The path distance doesn't take into consideration how deep in the WordNet hierarchy words can be found. Because the WordNet hierarchy increases in specificity the deeper it goes, the difference between concepts one node apart at the top of the hierarchy is much larger than differences further down. For example: the difference between 'invariability' and 'unchangeableness', which appear low on the hierarchy in the above diagram, is barely noticeable in English, but the concepts for 'people' and 'grouping', which appear near the top of the diagram, are very different.

For this reason, researchers have been working on alternative similarity metrics to path distance (Banerjee & Pedersen, 2003, Jurafsky & Martin, 2008; Ma, Fellbaum, & Cook, 2010; Resnik, 1995; Wu & Palmer, 1994). Many of these alternatives use the least common subsumer, which is the first hypernym that both synsets have in common. In the example above, the least common subsumer of 'monotony' and 'simpleness' is 'quality'.

Wu and Palmer (1994) use the ratio between the depth of the least common subsumer and the depth of the two synsets to determine similarity.

$$wpsim(syn1, syn2) = \frac{2(D_{lcs})}{D_{syn1} + D_{syn2}}$$

In the case of 'monotony' and 'simpleness', this evaluates to:

$$wpsim('monotony', 'simpleness') = \frac{2(3)}{7+4} \approx 0.55$$

In Wu-Palmer similarity, the closest two words can be in meaning is 1, in the case that they are in the same synset. Otherwise, this method ensures that the deeper into the hierarchy the least common subsumer is, the less the distance from the least common subsumer is important.

Starting with Resnik (1995), a group of new WordNet similarity metrics that use frequencies were developed. These similarity metrics use information content scores, based on the probability of any given word belonging to that concept. These scores assign a number to each concept that represents how much that concept contributes to the meaning of a word. For example, the lowest information content score belongs to 'entity'. Because every noun is a descendent of 'entity', the probability of a noun being an 'entity' is 1 and the information content score is 0. This makes intuitive sense; if every noun is an 'entity', then learning that a word represents an 'entity' gives us no information about its meaning.

These probability numbers are calculated over a corpus, ideally a word-disambiguated corpus, meaning that each word in the corpus is marked with which WordNet synset it is representing in the context of the sentence it appears in. In this case, each the appearance of a word in a specific synset counts towards the concept of the synset it belongs to and every synset above it in the hierarchy. In the absence of a word-disambiguated corpus, it is possible to split the count of each word evenly across multiple synsets. Dokusha takes this approach, and uses information content calculated over Japanese Wikipedia.

Dictionary-based similarity. ***D***ictionary based similarity metrics posit that the more closely related the dictionary glosses of two words are, the more similar the words are in meaning. Dokusha uses the English and Japanese definitions for the synsets found in WordNet to calculate the similarity of two words. The Extended Lesk algorithm (Banerjee & Pedersen, 2003) calculates a similarity score based on the number of overlaps in the glosses of the two words. The advantage to the dictionary based methods is that it can be extended in the future to evaluate the similarity of meaning between words that do not appear in WordNet, where the WordNet metrics fail if even one of the words cannot be found in the

WordNet database, as long as an additional dictionary is used, such as the Electronic Dictionary project, EDICT (Breen, 1993). For words that exist in WordNet, the addition of a gloss-based similarity metric can increase accuracy (Ma, Fellbaum, & Cook, 2010).

After the candidates are ranked according to word similarity, they are filtered by level. If no good candidate exists at or below the target level, candidates that are above the target level, but below the level of the original word are considered. The selected candidate is then conjugated into the same form as the original target word in the sentence, and is inserted into the sentence in place of the original word.

Combining the metrics. Different definitions of similarity produce different metrics for measurement, and a combination of these is needed to find an appropriate substitute. The vector similarity metric measures the similarity of context, while WordNet similarity measures similarity of meaning.

To illustrate the difference, we can look at what words are generated as synonym candidates for 江戸 *Edo*, the former name of Tokyo, and the seat of power for the Tokugawa shogunate during the Edo period. The best matches for Edo as measured by vector similarity are 鎌倉 *kamakura* and 平安 *heian,* both of which are names of other periods in Japanese history. The WordNet matches include 東京 *Tokyo* and 日本の首都 *Nihon-no-shuto* (the capital of Japan), both of which are alternate names or titles for the city of Edo.

Dokusha combines similarities calculated from WordNet and vector similarity. If the path between two words can not be found in WordNet, then the WordNet path similarity will not return any useful information. This may happen if the words are not of the same part of speech, for example. In this case, the glosses of the words are pulled from WordNet and the

Extended Lesk algorithm (Banerjee & Pedersen, 2003) is used instead. The score generated by Extended Lesk is on a different scale than the path similarity, so each candidate score is divided by the best score of the candidates then multiplied by 10, producing a number between 0 and 10, where lower numbers represent more similar words. Any given target word uses either the path distance, or the adjusted Lesk distance for all of its candidate words. The path distance is capped at 10, so the two scores are roughly equivalent.

The vector similarity is then added to this score. As many of the candidates had a cosine distance of 0.3 to 0.6 away from the target word, the vector similarity is multiplied by 30. This makes the range from 0 to 60, with most words falling in the 9 to 18 range. This matched the 0 to 10 range of the WordNet similarity best, so the two metrics would be given approximately equal weight.

The candidate score was then adjusted for word level. All candidates at a higher level than the original target word were eliminated, but candidates at a higher level than the target level were considered provided they were still at a lower level than the original word. Words at the target level of the simplified text were left alone, while candidates at a lower level than the target level were lightly penalized. Each level below the target level added five to the candidates score. This allowed candidates that were sufficiently closer to the target word in meaning and context to be chosen over others, even if the level of the candidate word was lower than the target level. Words above the target level had 10 per level added to the distance score, so only if there were no appropriate words at or below the target level would the higher levels be considered. This still allowed the overall level of the text to be reduced, even if hitting the target level was impossible.

Once all of the candidates have been scored, the candidate with the lowest score was selected for substitution. The winning candidate was then adjusted to the correct part of speech as necessary, and conjugated to match the context of the original target word.

Generating the Display Text

Finally, after the sentences have been broken down, grammatically simplified, and lexically simplified to meet or approach the target level, the simplified text is written to a file using HTML. HTML was chosen because it can be displayed easily on any device and supports ruby text. Ruby text is a way of displaying small text above normal text, rather than inline. This enables furigana to be displayed over kanji.

Dokusha adds furigana to any word that contains at least one kanji which is above the target level. In order to encourage readers to pay attention to the kanji characters themselves as well as its furigana reading, furigana only appears the first time the word appears in a paragraph. This means that a word with difficult kanji may appear without furigana, but only after its pronunciation has appeared once in furigana elsewhere on the screen.

In addition to furigana, difficult words contain links that when clicked reveal an English gloss of the word in the form of a popover, as shown in the image below.

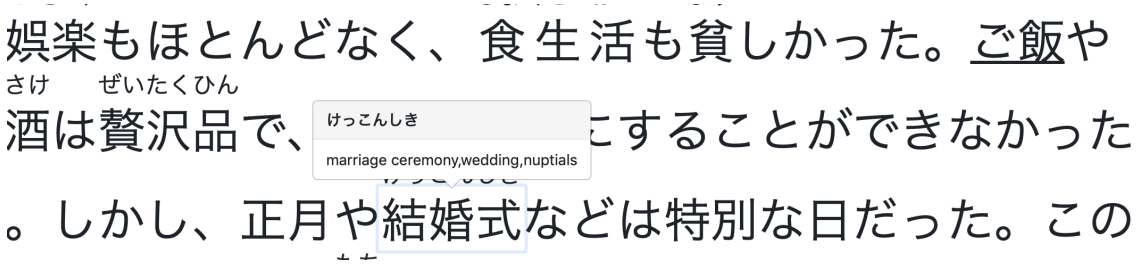

Image 1: Word glossing in Dokusha's output

Evaluation and Analysis of Dokusha

The goal of Dokusha is to generate simple, comprehensible text at an appropriate level for a second language learner. The evaluation of Dokusha aimed to answer the following two questions: 1) Does Dokusha generate grammatical and semantically correct Japanese text? 2) Does Dokusha simplify the text enough to be useful to a second language learner?

To address the first question, text simplified by Dokusha was given to native Japanese readers who were asked to correct it. The results, which will be examined in more detail in the following sections, showed that Dokusha made few grammatical errors, but a significant amount of semantic errors. To discover the source of these errors, each stage of lexical simplification was evaluated carefully.

Lexical Simplification

Lexical simplification is a multi-stage process including identification of complex words, generation of candidate words for substitution, and synonym selection. Evaluation of Dokusha's lexical simplification system aims to identify the successes and areas for improvement at every stage. To test the efficacy of Dokusha's lexical simplification system, three texts were randomly chosen from three different genres (textbook passages, newspaper articles, and book chapters) for evaluation. These texts were simplified by Dokusha to level N4, an advanced-beginning level of Japanese.

Dokusha's lexical substitutions across these three texts were classified into the following categories, based on Shardlow (2014a): problems with parsing, problems with

complexity identification, problems with substitution generation, or the substitution changing the meaning or grammar of the word used. Dokusha's errors break down as follows:

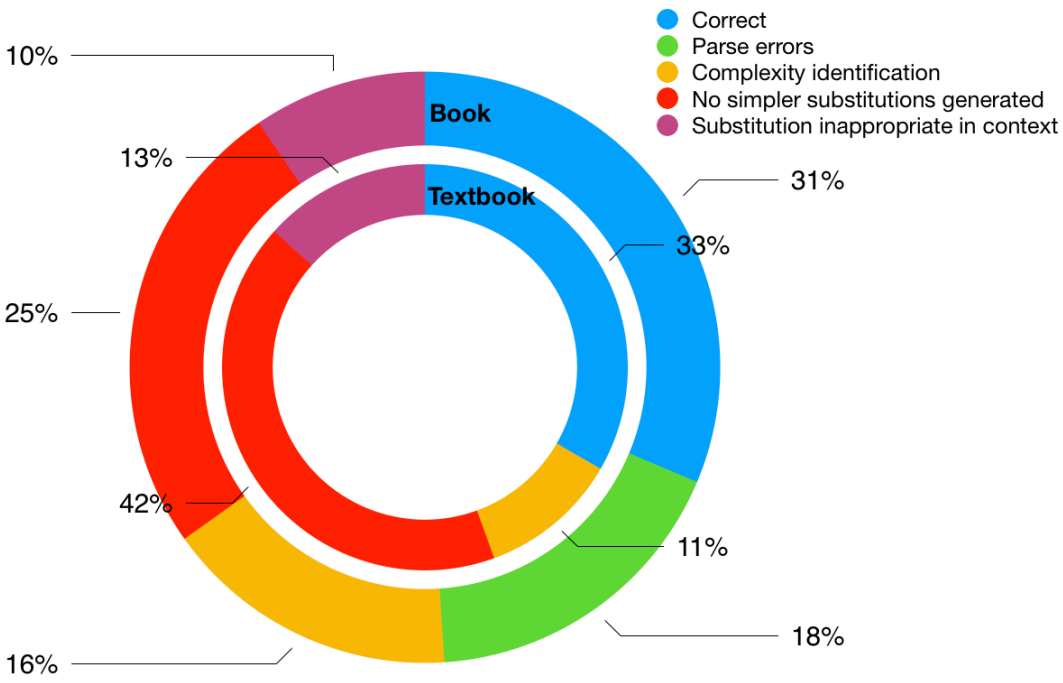

Figure 7: Error analysis of unique lexical substitutions
Note: News article breakdown not yet available.

The book had the most errors due to incorrect parses. This is because the KNP/Juman syntactic analysis tool used by Dokusha did not correctly parse the Kansai dialect, splitting many verb endings into small morphemes and mislabeling them as nouns, resulting in them inappropriately being marked for substitution.

Identification of complex words. The first stage of the lexical simplification process to be evaluated was whether or not Dokusha can correctly identify complex words that would be a good candidate for simplification. This question involves issues at both ends of the correction spectrum; if a word is not complex, but is identified by Dokusha as complex, then

it may be needlessly substituted, potentially robbing the text of some of its meaning along the way. This may also introduce the possibility of a grammatical error. On the other hand, if a word is complex, but is not identified by Dokusha as complex, then a complex word will remain in the text unsimplified, which will keep the text above the target level of simplification.

Finding the appropriate amount of correction needed to simplify a given text is crucial to the success of a test simplification system. Shardlow (2014a) found that 71% of the lexical simplification errors the analyzed system produced resulted either from words not being identified as complex (12%) or words being incorrectly identified as complex (59%).

The goal of Shardlow's system was to simplify the vocabulary of a text as much as possible, whereas the goal of Dokusha is to simplify the vocabulary to a specific reader's level. In the case of Dokusha, words that are likely to be unknown to a given reader should be considered for substitution. Defining the concept of an individual's reading level in terms that are both comprehensible to a computer and reasonable to execute requires some simplifying assumptions to be made. Dokusha approaches this problem as follows: Dokusha assumes that a reader at a given level of Japanese (as defined by the JLPT levels) will know all the words at and below that level, that all words written in katakana will be understandable to a native English reader, due to the vast majority of katakana words being loanwords borrowed from English, and that any word written using only known kanji would be readable.

As mentioned in Shardlow's (2014a) study, the decision about whether a word is complex enough to be substituted is a subjective one better suited to a human annotator than to an automated classification system. In an effort to reduce the subjectivity of this choice,

and to capture a case study of how effective Dokusha might be at reducing the level of a text for a specific reader, the decision was made to use a human annotator, a native English speaker at an advanced beginning level of Japanese. The annotator was given a list of all the words in all three texts and asked to mark whether or not they considered the word complex. This list was compared to the list of content words that Dokusha classified as complex: non-katakana words at level N1, N2, N3, or those not contained in the JLPT word lists, that are not proper nouns or auxiliary verbs.

	Textbook	Newspaper	Book	Overall
True Positives	23	101	132	254
False Positives	15	19	113	146
False Negatives	3	3	51	57
True Negatives	66	107	403	546
Precision	0.605	0.842	0.539	0.635
Recall	0.885	0.971	0.721	0.817
F Score	0.719	0.902	0.617	0.714

Table 1: Complex word classification

In classification tasks such as this one, it is common to look at precision and recall values to evaluate the systems effectiveness at classification. In Table 1 above, precision is the percentage of words that Dokusha marked as complex, and were also annotated as complex; this was calculated by dividing the count of true positives (the number of words both Dokusha and the human annotator identified as complex) by the total number of words Dokusha marked as complex. Recall is the percentage of words that the annotator marked as complex that Dokusha also recognized as complex; the count of true positives over the total number of words the annotator marked as complex. Like Shardlow's (2014a) system,

Dokusha had a lower precision rate than recall rate. That is, Dokusha tended towards rating too many words as complex across all three genres, marking more incorrectly as complex than incorrectly as simple.

The f-score is a standard calculation in classification tasks that balances the precision and recall scores, treating them as equally important. As some systems may have higher precision and others may have higher recall, the f-score provides a useful metric for comparing results side by side with a single score. Comparing the f-score across the textbook passage, newspaper article, and book chapter, it becomes clear that Dokusha had the most errors when classifying the words in the book, and was the most accurate with classifying difficult words in the newspaper article.

Of the three texts run through Dokusha's lexical simplification system, the complex word identification process had the most difficulty with the words in the book excerpt. The false positives from the book, the words that Dokusha identified as complex but the annotator did not, tended to be morphological variants of common words that occur more commonly in spoken Japanese than in written Japanese. The book chapter was the only one of the three passages tested to contain dialogue, which may have been a factor in the discrepancy. Dokusha also flagged many regional variants of common phrases as complex. The (randomly chosen) book excerpt was written in the Kansai dialect of Japanese. The Kansai dialect is not covered by the JLPT word list, which only covers standard Japanese. The annotator studied Japanese while living in the Kansai area of Japan, which may account for the large portion of false positives in the book excerpt list.

Of the 51 false negatives, words the annotator felt were complex, but that Dokusha did not mark as complex, 35 were written in katakana. This indicates that the assumption

that words written in katakana would be easy to understand for a native English speaker was overly simplistic. The majority of the katakana words in the book excerpt did not come from English, but were orthographical variants of Japanese words which had been written in katakana for emphasis, making the annotator's knowledge of English unhelpful in deciphering those words. This is not common outside of dialogue in works of fiction, and this particular book passage was entirely dialogue, making this problem likely to be specific to dialogue-heavy books.

The book's complexity identification errors were mostly due to its inability to identify variants of words contained in the JLPT word lists. 39 of the 46 complexity identification errors were morphologic or orthographic variants of words in the JLPT lists at level N4 or N5, and were misclassified as complex. These already low level words often failed to generate simpler substitutions. When the orthographic variants of low level JLPT words did generate a simpler substitution, they were often simply the same word in the orthography on the JLPT list.

	Textbook	Newspaper	Book	All
Not in JLPT	20.73%	23.49%	28.20%	26.76%
N1	0.00%	1.34%	0.56%	0.67%
N2	6.10%	20.13%	3.57%	7.19%
N3	19.51%	34.90%	13.72%	18.51%
N4	24.39%	4.70%	9.96%	10.25%
N5	24.39%	14.77%	30.08%	26.10%
Katakana	4.88%	0.67%	13.91%	10.52%

Table 2: Breakdown of JLPT leveled words

From the breakdown of where and how the words were classified into the JLPT levels, shown in table 2 above, it is easy to see that the book used more katakana words than

the other two genres. Nearly 14% of the unique words in the book text were written in katakana. Additionally, the book contained the highest percentage of words not in the JLPT lists.

The JLPT breakdown across all three passages reveals a flaw in the current leveling system: The JLPT lists are not comprehensive enough. Over a quarter of the unique words encountered in these texts were not in the JLPT lists.

Table 3 below shows the breakdown of the annotator's knowledge of the words in each passage broken down by JLPT level. While N5 and N4 words were mostly known, there is a clear drop-off at N3. The annotator was unfamiliar with any N1 words that appeared in the texts. However, the annotator was familiar with many of the words which were not in the JLPT lists, further emphasizing the need for more comprehensive lists, or a better classification technique for JLPT levels.

	Textbook	Newspaper	Book
Not in JLPT	47.06%	14.29%	43.33%
N1	N/A	0.00%	0.00%
N2	20.00%	10.00%	26.32%
N3	37.50%	20.75%	58.90%
N4	90.00%	71.43%	88.68%
N5	100.00%	95.45%	93.13%
Katakana	75.00%	100.00%	54.05%

Table 3: Annotated word complexity by JLPT level

Table 3 shows that, despite the oversimplification in the assumption that a reader at a given level will know all the words at and below the level, and no words above the level, the JLPT levels may actually be a decent proxy for word knowledge. Ideally, Dokusha could know precisely which words are in a given reader's sight vocabulary, which words the reader

has been exposed to before, and which words the reader has never encountered before. With such fine grained information, Dokusha would have a higher accuracy when choosing words to substitute. Until this point, at least for our case-study learner, the JLPT levels seem like a close enough approximation of word knowledge.

Dokusha has a tendency to be aggressive with marking words as complex, due to a simple lack of data on JLPT word levels and some faults in the recognition of morphologic and orthographic variants of words in the JLPT lists. The few words that are complex, but not flagged as such by Dokusha are due to an oversimplification in the classification of katakana words.

Substitution generation. Dokusha generates a list of the top 10 candidates it finds for each target “complex” word, irrespective of whether or not they are deemed simpler or more complex than the target word. To test the acceptability of the selection candidates, a human reader hand-marked the acceptability of each candidate for substitution in the sentence. In this context, 'acceptable' was defined as being substitutable for the original word in the sentence, without changing the overall meaning or grammar of the passage. For example, in the sentence "Until the Edo period, people's lives were very monotonous," 'Kamakura period' was judged an inappropriate substitution for 'Edo period', as the meaning of the passage would change. In the sentence, "Days such as New Year's day, festival days and wedding days were special", however, 'celebration' was judged an acceptable substitution for 'wedding'.

	Textbook	News	Book	Overall
% of words with acceptable substitutions	74.19%	67.83%	63.85%	66.67%
% of words with acceptable simpler substitutions	48.39%	22.45%	26.15%	27.41%

Table 4: Percentages of target words with acceptable candidates for substitution using word vectors

Table 4 above shows the results from the candidate generation step using the word vector method as explained in the implementation section. The top ten most similar words by cosine similarity were carefully considered for appropriateness of meaning. While acceptable substitution candidates were found for around two-thirds of the target words, more than half of these candidates were above N4, the target level for simplification. Just over 27% of the target words were able to generate a suitable candidate at an appropriate level. Especially for the lowest-level (N3) target words, very few candidates with a lower level (N4 or N5) were generated at all.

An additional problem is that just over 20% of the candidates generated were simply morphological variants of the original target word, and were therefore not deemed acceptable substitutions. For these reasons, the number of acceptable substitution candidates generated was quite low.

Generating candidates from WordNet synonyms and hypernyms does not generate as many acceptable candidates as the word vector method. This is partly because many of the target words were not found in WordNet. Despite this, WordNet generates substitution

candidates that are simpler than the target word disproportionately more often than the word vector method.

	Wordnet	Word Vectors	Combined
% of words with acceptable substitutions	33.93%	66.67%	73.81%
% of words with acceptable simpler substitutions	23.81%	26.49%	42.86%

Table 5: Percentages of target words with acceptable candidates for substitution across all texts

As shown in Table 5 above, using a combined method, where the synonym candidate lists generated through WordNet and the word vectors are combined, increases coverage. The number of words with acceptable substitutions at levels N4 and N5 increases noticeably. Also, the WordNet and word vectors methods generate acceptable simpler substitutions for almost entirely different sets of words.

Synonyms in context. The final test of the lexical simplification system is to see if the top rated synonyms would be acceptable in the original context of the target word. Of the 40 words that were marked as "complex" in the textbook passage, two words failed to generate any substitutions at all, and an additional 17 words failed to generate any simpler substitutions. The final 21 words were substituted by Dokusha with their simplified versions to form the lexically simplified text.

The lexically simplified text was sent to two native readers of Japanese, neither of whom were given the original text, and who were given no indication of what words may

have changed between the two versions. They were asked to highlight any words that felt out of place or confusing.

Among those 21 simplified words, five were marked by the native readers as incorrect words, and an additional one word was marked as a correct word in an incorrect form. The other 15 substitutions were unnoticeable to the native reader. Interestingly, several words from the original text were marked as incorrect by the reader, many of which were in sentences that were left entirely unchanged.

In the book chapter, of the 116 words for which a simpler substitution candidate was found, 27 were marked as incorrect by the native readers. When asked to correct the sentences, the native reader changed the grammatical form of one word, removed one word, and suggested alternative words for the other 25 incorrect substitutions. This suggests that Dokusha's mistakes are largely semantic, not syntactic. Interestingly, the substitutions suggested by the native readers for four of the incorrect substitutions were the original target word that had been substituted out. The readers had not been exposed to the original text, so this seems to indicate that those four words are not substitutable in this context.

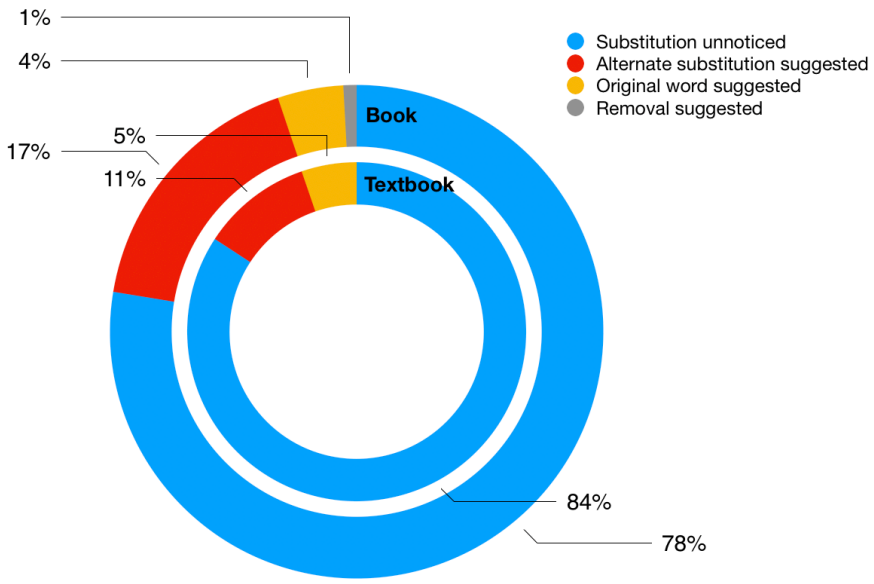

Figure 8: Synonyms deemed acceptable in context by native readers

Evaluation Using Existing Readability Scoring Systems

After evaluating Dokusha's ability to produce acceptable, grammatical Japanese, it remains to evaluate how well the system meets the goal of increasing readability. To this end, Dokusha was evaluated by an online automatic readability calculator (Hasabe & Lee, 2015), and compared against the vocabulary coverage goal of 98%.

Using Hasabe and Lee's (2015) system, the original texts were classified as follows: upper elementary (book chapter), lower intermediate (textbook passage) and upper advanced (news article). Given the large number of kanji and words of Chinese origin, it is unsurprising that the news article was classified as upper advanced. After running the article through Dokusha, the adjusted text was classified as lower advanced. The number of words of Chinese origin in the first text (331) was reduced to 306 in the adjusted text, lowering the number of kanji in the article from 599 to 572. The readability score produced by Hasabe and Lee's (2015) system increased from 1.28 to 1.55.

The classification of the book chapter as the easiest of the three passages highlights the shortcomings of readability scoring systems. The book started at a high readability score of 5.02, based partly on the large number of words of Japanese origin written in hiragana. This is slightly misleading. Most of the words in the book were in the Kansai dialect and nearly 30% of the unique content words were not included in any of the JLPT lists, making this arguably the hardest of the three texts for a language learner to comprehend. 294 words were considered too difficult by Dokusha and were marked for substitution. Some of these words were switched out for common words of Chinese origin that contain simple kanji. This means that the number of words of Chinese origin increased slightly in the book from 183 to 188, increasing the number of kanji from 651 to 715, and decreasing the hiragana count from 2627 to 2479. Despite this, Dokusha increased the book chapter's readability score very slightly from 5.02 to 5.03.

Automated systems of judging readability are not always reliable in the process of text simplification. These systems rely on a complex assessment of different sentential features such as sentence length and syllable count, assuming that shorter sentences and words with fewer syllables make a text easier to read. It is possible to 'game the system' by tailoring the text simplification algorithm to the readability assessment measure; if the assessment relies heavily on sentence length, simply cutting words out of the sentence will produce a higher readability score, but might make the sentence difficult to understand, or drastically change its meaning. For Japanese, this is further complicated by the issue of orthography. Automated systems do not take into account the presence or absence of furigana, the phonetic text above kanji. Also, the number and complexity of kanji are two features of sentences often used to judge readability, while Kanji knowledge varies widely

among readers and Kanji words written in hiragana may actually be more difficult to read for students at some levels (Harada, 1988; Koda, 1992).

Despite its faults, it is interesting to see how an outside automated system would assess the difficulty level of the passages before and after they had been run through Dokusha's simplification. Hasebe and Lee (2015) created a system that uses features of the text, such as number of words per sentence, and the proportion of words of Japanese origin and Chinese origin, to estimate readability. As mentioned above, the words of Chinese origin are predominately written in kanji and are considered more complicated words than those that originated in Japan.

Coverage Percentages

As mentioned in a previous section, for English the target percentage of known words for comprehensibility is 98%. There have not been enough studies to check that this number also holds true for Japanese language learners, but it remains logical that the higher the coverage can be, the more comprehensible a text will be for readers.

Consistent with previous methods of evaluating the complexity of potential word substitutions, the following coverage percentages have been calculated with the assumption that a reader at the N4 level will know all the words at N4 and N5, and none of the words at the levels N1-N3 or not included in the JLPT levels.

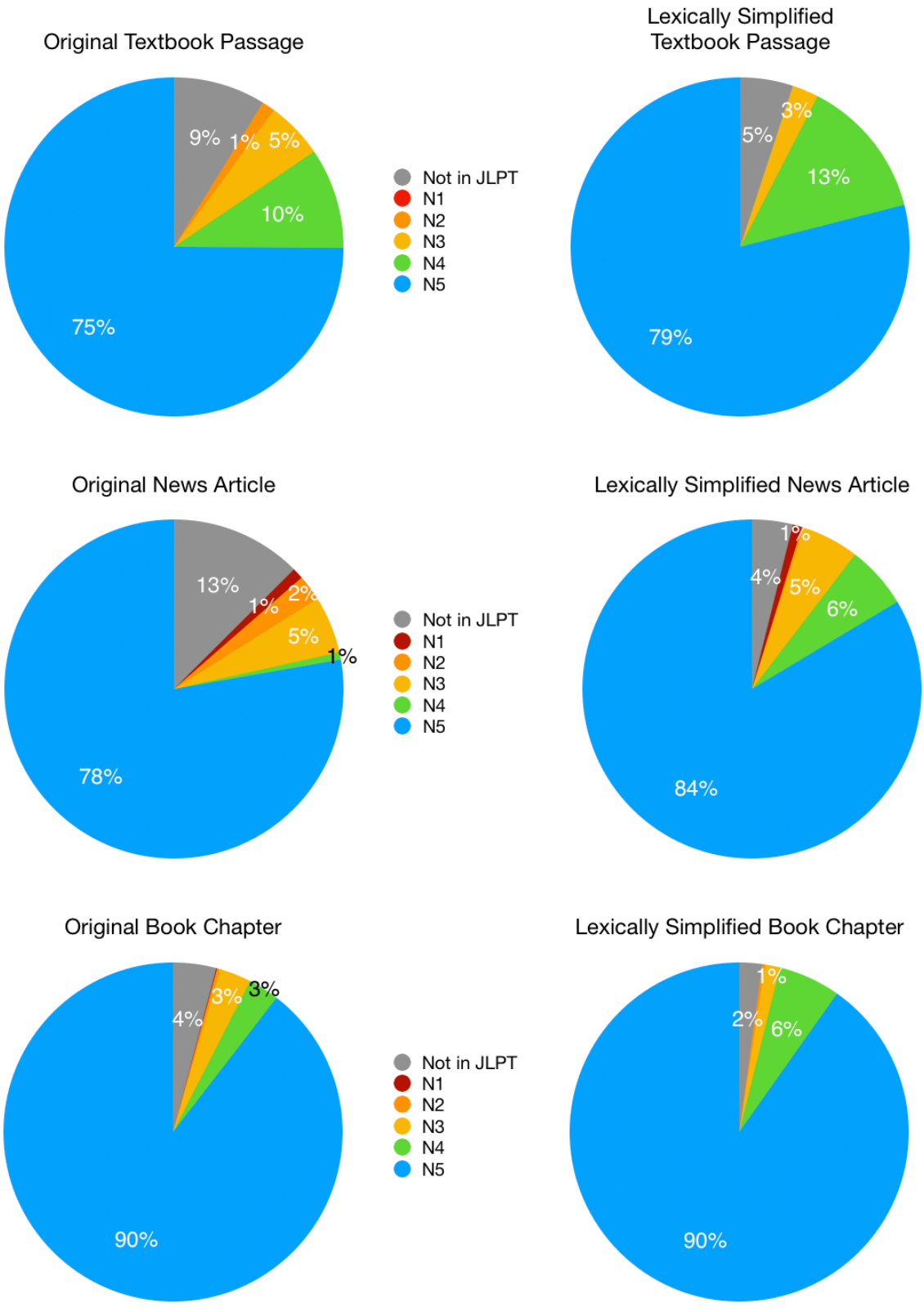

Figure 9: JLPT coverage percentages

Figure 9 shows how Dokusha classifies all word tokens, before and after simplification. Given the large number of words not in the JLPT, and the fair likelihood that a word at the N3 level will be familiar to a N4 reader, adjusted numbers based on the manual complexity analysis were calculated to estimate the percentage of words probably known to an N4 reader. As shown in Figure 10, adjusting for the number of words probably known to an N4 reader puts both the adjusted newspaper article and the adjusted book chapter past the 98% goal, and the significant improvement in the textbook passage is made apparent.

Over 30% of complex words are still being successfully substituted for easier vocabulary words, despite errors in parsing, complexity identification, and substitution generation. These substitutions are reducing the overall lexical load of the text. While it was not conclusively demonstrated that the text has reached the coverage needed for extensive reading at the N4 level, it remains that Dokusha can make text more approachable for learners at an intermediate level of Japanese.

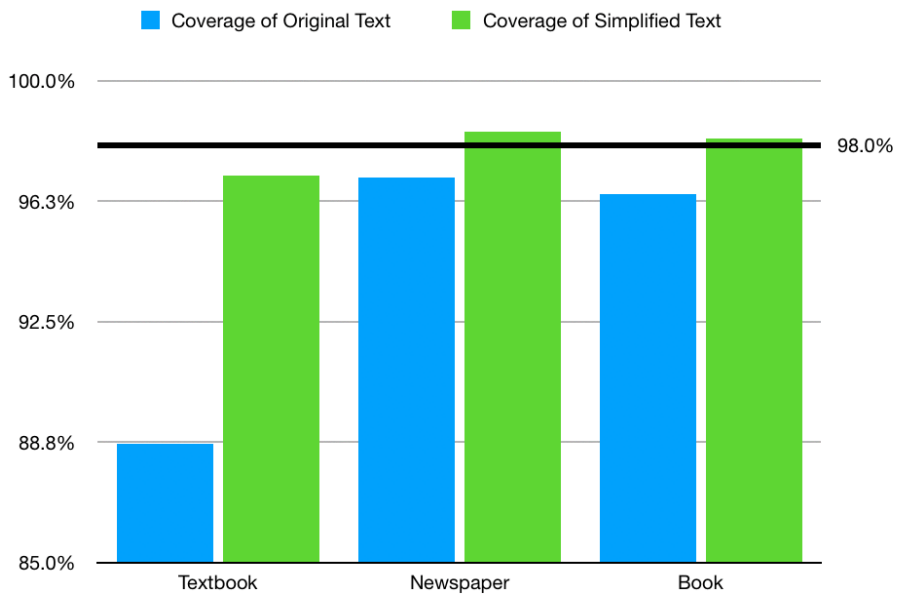

Figure 10: Adjusted word knowledge coverage percentages

Future Work

While Dokusha was able to reduce the level of the text it was evaluated on, it had significant shortcomings that should be addressed in future work on the project. Its lexical substitution system is only capable of swapping out one word for one word, and cannot handle multi-word substitutions or simply removing unneeded words. In the case of the textbook passage, for example, the majority of errors came from difficult words having no level appropriate synonyms generated. Some of these words were not necessary for the meaning of the overall sentence and could have been removed altogether. The news article used many complicated words that were specific enough to not have level appropriate substitutions. Enabling multi-word substitutions could have reduced the lexical load, even as the length of the text increased, by swapping individual words out for explanatory phrases (Drandarević & Saggion, 2012). Another option for handling domain specific words would be the addition of inline definitions. This would make the text more educational, actually teaching the meaning of new words rather than eliminating them, although the change in readability would be difficult to quantify (Shardlow, 2014b).

The current system of lexical simplification also fails to take into account the context of the target word when selecting candidates. This is one of the leading causes of some of the substituted words sounding odd to the native readers. In the book, there were multiple instances where one word was substituted by the same word in two different contexts, one of which was correct, and the other was incorrect. While Dokusha selects words that may be substitutable in some situations, it may not be the case that the target word can be substituted in the context of the particular sentence.

Dokusha's syntactic analysis performs poorly on dialect and colloquial Japanese. The word vectors, trained on Japanese wikipedia, also are biased toward the standard written form of Japanese. Dialects in particular present an interesting challenge for the design and implementation of Dokusha. Learners generally study standard Japanese, so regional words do not appear on the JLPT. These words are all marked by Dokusha as needing replacement. However, Japanese as a foreign language students living outside of Tokyo are exposed to dialects regularly, and may want the exposure to more region-specific vocabulary. Removing colloquial or regional expressions also removes characterization of characters in books, as often the particular style of speech is meant to convey a certain trope in fiction. Rather than lexically substituting out non-standard Japanese, it may be more interesting to convert to standard Japanese in the pop-up glosses. More work on dialect detection and conversion would be needed to realize such a feature.

More research is needed in Japanese language education to understand the role of furigana in vocabulary and kanji acquisition as well as in text readability. The combined information of the ideographic character accompanied by its pronunciation, could act as additional information which, along with the context of the sentence, could make unknown words more guessable. Using kanji increases a reader's understanding of word boundaries, and more exposure to the characters increases retention. Despite its potential benefits, furigana is an incredibly understudied tool in Japanese as a foreign language education. The readability algorithms don't take the presence or absence of furigana into account in their calculations, but many Japanese language learners do when evaluating resources for potential study. Not only does furigana have potential to be a useful tool for Japanese language learners, it is frequently found in authentic materials written for a native Japanese audience.

More studies need to be conducted on how the use of furigana effects Japanese literacy for second language readers.

While English literacy for both first and second language readers is a well researched field, the field of study concerning Japanese literacy, especially for second language readers, is lagging behind. Among other data-driven insights, it would be helpful to have a clear coverage percentage, such as the 98% goal, for Japanese.

Conclusion

While there is plenty of room for further improvements to the system, such as a more robust parser to handle colloquial and regional Japanese, and to better classify non-JLPT words into the JLPT categories, the results presented validate the potential for simplification systems such as Dokusha to significantly improve the readability of texts for learners of Japanese. With further investment in comprehensive word difficulty ranking, and more nuanced handling of alternative orthographies like katakana, Dokusha could be a very effective tool for Japanese language learners.

Given the growing community of independent Japanese language learners who rely on reading as a way to learn autonomously, and the increasing popularity of extensive reading programs as a method of strengthening language skills in the classroom, the need for a large body of Japanese language learner literature is apparent. Writing graded readers is a time consuming process, so the quantity of learner targeted text is increasing slowly. Using text simplification methods, Dokusha demonstrates the viability of a system to automatically reduce the level of text specifically for second language learners. With some modifications, this system could present learners with model text at a targeted level for their personal

language growth, generated from texts chosen by the learners themselves, which in turn increases the likelihood that the text content will be relevant and motivating to the reader.

References

Baeza-Yates, R., Rello, L., & Dembowski, J. (2015). CASSA: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 NAACL,* 1380-1385.

Bell, T. (2001). Extensive reading: speed and comprehension. *Reading Matrix: An International Online Journal,* 1(1).

Besner, D. (1990). Orthographies and their phonologies: A hypothesis. *Bulletin of the Psychonomic Society*, 28(5), 395-396.

Biran, O., Brody, S., & Elhadad, N. (2011). Putting it simply: a context-aware approach to lexical simplification. *In Proceedings of the 49th ACL*, 496–501.

Breen, J.W. (1993). *A Japanese Electronic Dictionary Project (Part 1: The Dictionary Files)*. (Technical Report). Department of Robotics & Digital Technology, Monash University.

Brimo, D., Apel, K., & Fountain, T. (2017). Examining the contributions of syntactic awareness and syntactic knowledge to reading comprehension. *Journal of Research in Reading*, 40(1), 57–74.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G.A. Miller (Eds.), *Linguistic theory and psychological reality* (264-293). Cambridge, MA: MIT Press.

Chandrasekar, R. & Srinivas, B. (1997). Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3), 183 – 190.

Clarke, M. (1980). The short circuit hypothesis of ESL reading -- or when language competence interferes with reading performance. *Modern Language Journal,* 64(2), 203-209.

Cummins, J. (1979). Linguistic Interdependence and the Educational Development of

Bilingual Children. *Review of Educational Research*, 49, 222-251.

Day, R., & Bamford, J. (1998). *Extensive Reading in the Second Language Classroom*. Cambridge: Cambridge University Press.

Day, R., & Bamford, J. (2002). Top ten principals for teaching extensive reading. *Reading in a Foreign Language,* 14(2), 136-141.

De Belder, J., & Moens, M.-F. (2010). Text simplification for children. In *Proceedings of the 2010 SIGIR Workshop on Accessible Search Systems,* 19-26.

Elley, W. & Mangubhai, F (1983). The impact of reading on second language learning. *Reading Research Quarterly, 19(1),* 53-67.

Elley, W. (2000). The potential of book floods for raising literacy levels. *International Review of Education,* 46(3/4), 233-255.

Eskey, D. E. (2002). Reading and the teaching of L2 reading. *TESOL Journal,* 11, 5-9.

Everson, M. E., & Kuriya, Y. (1998). An Exploratory Study into the Reading Strategies of Learners of Japanese as a Foreign Language. *Journal of the Association of Teachers of Japanese*, 32(1), 1–21.

Farrell, T. (2009). *Teaching Reading to English Language Learners: A Reflexive Guide*. Thousand Oaks, CA: Corwin Press.

Glavaš, G., & Štajner, S. (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora? *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)*. 63-68.

Google. (n.d.) Google Translate. Retrieved from https://translate.google.com

Grabe, W., & Stoller, F. L. (2013). *Teaching and Researching: Reading* (2nd ed.). New York,

NY: Routledge.

Harada, F K. (1988). The effect of three different orthographical presentations of a text upon the reading behaviors of native and non-native readers of Japanese: An eye-tracking study. Doctoral dissertation, Ohio State University.

Hasebe, Y., & Lee, J. (2015). Introducing a readability evaluation system for Japanese language education. *Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese*.

Hauptman, P., Wesche, M., & Ready, D. (1988). Second-language acquisition through subject-matter learning: A follow-up study at the University of Ottawa. *Language Learning*, 38(3), 433-475.

Hayashi, Y. (1992). A three-level revision model for improving Japanese bad-styled expressions. *COLING-92*, 665-671.

Hayes, D., & Ahrens, M. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of Child Language*, 15(2), 395-410.

Hitosugi, C., & Day, R. (2004). Extensive reading in Japanese. *Reading in a Foreign Language*. 16 (1), 20-39.

Hirsh, D, & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2), 689-696.

Horiba, Y. (1990). Narrative comprehension processes: a study of native and non-native readers of Japanese. *Modern Language Journal*, 74(2), 188–202.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review,* 61(3), 355-382.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond *A Clockwork Orange*: Acquiring second

language vocabulary through reading. *Reading in a Foreign Language,* 11(2), 207-223.

Hyland, F. (2004). Learning autonomously: Contextualising out-of-class English language learning. *Language Awareness*, 13(3), 180-202.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008). Development of the Japanese WordNet. *LREC 2008*. 2420-2423.

Izumi, S. (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale of the output hypothesis. *Applied Linguistics,* 24(2), 168-196.

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing (2nd ed.)*. Upper Saddle River, NJ: Prentice Hall.

Kaiho, H. (1975). Kanji imi jouhou chuushutsu katei. [The kanji meaning information abstraction process]. *Journal of gakugei Tokushima University. Social science,* 24, 1-7.

Kajiwara, T., Matsumoto, H., & Yamamoto, K. (2013). Selecting proper lexical paraphrase for children. *Proceedings of the 25th ROCLING*, 59–73.

Kirwan, L. (2003). The role of furigana in Japanese script for second language learners of Japanese. Doctoral dissertation, The University of Queensland.

Klerke, S. & Sogaard, A. (2013). Simple, readable sub-sentences. *Proceedings of the ACL Student Research Workshop,* 142–149.

Koda, K. (1992). The effects of lower-level processing skills on FL reading performance: Implications for instruction. *The Modern Language Journal*, 76(4), 502-512.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second

language reading development. *Language Learning*, 57, 1-44.

Koda, K. (2008). Impacts of prior literacy experience on second- language learning to read. In K. Koda & A. Zehler (Eds.), *Learning to Read Across Languages: Cross-Linguistic Relationships in First- and Second-Language Literacy Development* (68-96). New York, NY: Routledge.

Kondo−Brown, K. (2006), How Do English L1 Learners of Advanced Japanese Infer Unknown Kanji Words in Authentic Texts?. *Language Learning*, 56: 109-153.

Krashen, S. (1985). *The input hypothesis*. London: Longman

Krashen, S. (2003). *Explorations in Language Acquisition and Use*. New Hampshire: Heinemann.

Kurohashi, S. (1998). *Japanese Dependency/Case Structure Analyzer KNP version 2.0b6*. Kyoto University.

Kurohashi, S., & Nagao, M. (1994). A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4), 507-534.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (316–323). Clevedon: Multilingual Matters.

Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension, *Reading in a Foreign Language*, 22(1), 15-30.

Laufer, B., & Sim, D. (1985). Taking the easy way out: Non-use and misuse of clues in EFL reading. *English Teaching Forum*, 23(2), 7-10.

Lee, H., & Lee, J. H. (2015). The effects of electronic glossing types on foreign language vocabulary learning: different types of format and glossary information. *Asia-Pacific Education Researcher*, 24(4), 591–601.

Lee, J., & Schallert, D. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *TESOL Quarterly*, 31(4), 713-739.

Leung, C. (2002). Extensive reading and language learning: A diary study of a beginning learner of Japanese. *Reading in a Foreign Language*, 14(1), 68-81.

Light, T. (1970). The reading-comprehension passage and a comprehensive reading programme. *ELT Journal*, 24(2), 120–124.

Lightbown, P., & Spada, N. (2013). *How Languages are Learned (4th edition)*. Oxford, UK: Oxford University Press.

Ma, X., Fellbaum, C., & Cook, P. (2010). A multimodal vocabulary for augmentative and alternative communication from sound/image label datasets. *Proceedings of the NAACL Human Language Technologies (HLT 2010) Workshop of Speech and Language Processing for Assistive Technologies*, 62-70.

Mason, B., & Krashen, S. (1997). Can extensive reading help unmotivated students of EFL improve? *I.T.L. Review of Applied Linguistics, 177-118*, 79-84.

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research,* 53(2), 253-279.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*.

Miller, G. (1995). WordNet: A lexical database for English. *Communications of the ACM,*

38(11), 39-41.

Mino, H., & Tanaka, H. (2011). Housou nyuusu no doushiren'youkeimeishi no heiika [Simplification of nominalized continuative verbs in broadcast news]. *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing,* 744-747.

Mori, Y. (1998). Effects of first language and phonological accessibility on Kanji recognition. *Modern Language Journal*, 82(1), 69.

Mori, Y., & Nagy, W. (1999). Integration of information from context and word elements in interpreting novel kanji compounds. *Reading Research Quarterly*, 34(1), 80.

Muljani, D., Koda, K., & Moates, D. (1998). The development of word recognition in a second language. *Applied Psycholinguistics*, 19(1), 99-113.

Muto, H. (2015). The effects of linearity on sentence comprehension in oral and silent reading. *Japanese Psychological Research*, 57(3), 194–205.

Nation, I.S.P. (2013). *Learning Vocabulary in Another Language*. NY: Cambridge University Press.

Nunes, B. P., Kawase, R., Siehndel, P., Casanova, M. A., & Dietze, S. (2013). As Simple as It Gets - A Sentence Simplifier for Different Learning Levels and Contexts, *IEEE 13th International Conference on Advanced Learning Technologies*, 128-132.

Paetzold, G. H. & Specia, L. (2013). Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. 116–125.

Paetzold, G. H. & Specia, L. (2015). Lexenstein: A framework for lexical simplification. In *Proceedings of the 53rd ACL*, 85-90.

Paetzold, G. H., & Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research,* 60, 549-593.

Pellicer-Sanchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do *Things Fall Apart? Reading in a Foreign Language,* 22, 31-55.

Perfetti, C. & Dunlap, S. (2008). Learning to read: General principles and writing system variations. In K. Koda & A. Zehler (Eds.), *Learning to Read Across Languages: Cross-Linguistic Relationships in First- and Second-Language Literacy Development* (13-39). New York, NY: Routledge.

Pickard, N. (1995). Out-of class language learning strategies: Three case studies. *Language Learning Journal*, 12, 35-37.

Pitts, M., White, H., & Krashen, S. (1989). Acquiring second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language*, 5(2), 271-75.

Pozzan, L, & Trueswell, J. (2016). Second language processing and revision of garden-path sentences: a visual word study. *Bilingualism*, 19(3), 636-643.

Pulido, D., & Hambrick, D. (2008). The virtuous circle: Modeling individual differences in L2 reading and vocabulary development. Reading in a Foreign Language, 20(2), 164-190.

Saragi, T., Nation, I.S.P., & Meister, C.F. (1987). Vocabulary learning and reading. *System*, 6(2), 72-8.

Sakuma, N., Sasanuma, S., Tatsumi, I., & Masaki, S. (1998). Orthography and phonology in reading Japanese kanji words: Evidence from the semantic decision task with homophones. *Memory & Cognition*, 26(1), 75-87.

Sasaki, H., & Matsumoto, N. (2010a). *Nihongo Nouryoku Shiken Taisaku Nihongo So-Matome N1 Bunpou [Japanese Language Proficiency Test Summary: N1 Grammar]*. Japan: Ask Publishing Co.,Ltd.

Sasaki, H., & Matsumoto, N. (2010b). *Nihongo Nouryoku Shiken Taisaku Nihongo So-Matome N2 Bunpou [Japanese Language Proficiency Test Summary: N2 Grammar]*. Japan: Ask Publishing Co.,Ltd.

Sasaki, H., & Matsumoto, N. (2010c). *Nihongo Nouryoku Shiken Taisaku Nihongo So-Matome N3 Bunpou [Japanese Language Proficiency Test Summary: N3 Grammar]*. Japan: Ask Publishing Co.,Ltd.

Sasaki, H., & Matsumoto, N. (2010d). *Nihongo Nouryoku Shiken Taisaku Nihongo So-Matome N4 Bunpou, Dokkai, Choukai [Japanese Language Proficiency Test Summary: N4 Grammar, Reading Comprehension, Listening Comprehension]*. Japan: Ask Publishing Co.,Ltd.

Sato, T., Matsunuma, M. & Suzuki, A. (2013). Enhancement of automatization through vocabulary learning using CALL: can prompt language processing lead to better comprehension in L2 reading? *ReCALL, 25(1)*, 143-158.

Scott, V., & de la Fuente, M. J. (2008). What's the problem? L2 learners' use of the L1 during consciousness−raising, form−focused tasks. *Modern Language Journal,* 92(1), 100-113.

Senoo, Y., & Yonemoto, K. (2014). Vocabulary learning through extensive reading: A case study. *The Canadian Journal of Applied Linguistics*, 17(2), 1-22.

Sergent, W. K., & Everson, M. E. (1992). The effects of frequency and density on character recognition speed and accuracy by elementary and advanced L2 readers of Chinese.

Journal of the Chinese Language Teachers Association, 27(1–2), 29–44.

Shardlow, M. (2014a). Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014),* 1583-1590.

Shardlow, M. (2014b). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing 2004*. 58-70.

Shiotsu, T, & Weir, C. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24(1), 99–128.

Stanovich, K. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, 16(1), 32-71.

Swaffer, J., & Woodruff, M. (1978). Language for comprehension: Focus on reading. *Modern Language Journal*, 62, 27-32.

Taylor, I. & Taylor, M. (1983). *The Psychology of Reading*. New York: Academic Press.

The Japan Foundation. (2018). Nihongo nouryoku shiken JLPT [The Japanese Language Proficiency Test JLPT]. Retreived from https://www.jlpt.jp.

Tsuchiya, M., & Sato, S. (2003). Automatic detection of grammar elements that decrease readability. *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*.

Umemura, C. (1981). Kana to kanji no moji-kinou no sai ni tsuite [Functional properties of Japanese letters (kana and kanji) in memory studies]. *Japanese Journal of Educational Psychology,* 29(2), 123-131.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, 15(2), 130-163.

Webb, S. & Macalister, J. (2013) Is text written for children useful for L2 extensive reading? *TESOL Quarterly*, 47 (2), 300-322.

Webb, S., & Rodgers, M. (2009). The lexical coverage of movies. *Applied Linguistics,* 30(3), 407–427.

Yamashita, J. (2002). Mutual compensation between L1 reading ability and L2 language proficiency in L2 reading comprehension. *Journal of Research in Reading, 25*: 81-95.

Yokoyama, S., & Imai, M. (1989). Kanji to kana no houki-keitai no sai ga tango no guuhatsu-kioku ni oyobosu kouka [The effect of orthographic difference between kanji and kana words on incidental memory]. *The Japanese Journal of Psychology,* 60(1), 61-63.