

7 Readability measurement of Japanese texts based on levelled corpora

LEE Jae-ho

Waseda University

HASEBE Yoichiro

Doshisha University

Abstract

A method is presented here to measure the readability of Japanese texts using levelled corpora. Two sets of levelled corpora were constructed for this purpose: one was used as model data to devise a readability measurement formula, and the other as test data to check the validity and reliability of the formula. Six-level model corpora were constructed at first using texts extracted from Japanese textbooks and Japanese Diet meeting transcripts. We examined these corpora both manually and statistically. Then a multiple regression analysis of the results of these examinations was carried out. Among the five models produced, the best model was selected and used to construct a readability formula. The formula was tested using the other set of levelled corpora based on 25 years of reading passages from the Japanese-Language Proficiency Test (JLPT), and its reliability was confirmed. A web-based system was also developed using the formula to aid teachers of Japanese in preparing reading materials that match student levels. The system also has much reading-related functionality, making it helpful to teachers and learners, as well as allowing wide access of the present research to a broad range of people involved in teaching, learning, and studying Japanese.

Keywords: readability, levelled corpora, regression analysis, web-based system

1 Background and purpose

Text readability studies aim to devise methods to measure reading difficulty in natural language texts. Research in this field has developed systematic procedures that rank the level of a given text based on various indices such as the mean number of words per sentence. There is a long tradition of such attempts for texts written in English, and a number of methods and formulae have been proposed (e.g., Flesch 1948; Smith and Kincaid 1970). In recent years, readability studies have also been actively pursued to measure texts in Japanese (e.g., Sakamoto 1964; Tateishi et al. 1988; Shibasaki and Hara 2010; Sakai 2011; Sato 2011). Moreover, several web-based systems targeted at Japanese native speakers have been developed utilizing various methods and formulae¹.

1 Shibasaki and Hara (2010) have made their online system available at the following website: <http://readability.nagaokaut.ac.jp/readability>, and Sato (2011) at: <http://kotoba.nuee.nagoya-u.ac.jp/sc/readability/index.html>.

No matter what the target language, virtually all studies in text readability measurements have been completed with the following two points in mind: 1) What are the essential factors that determine the level of the text? 2) How is it possible to formalize the relationship among various factors and produce a readability formula? As to question 1, the factors need to be broadly divided into two types. On the one hand, there are macro factors such as topics and coherence, and on the other, there are micro factors such as levels of vocabulary items, degrees of complexity of grammatical structures, and length of words and sentences. Focused primarily on the factors of the latter type, Shibasaki and Hara (2010) produced a readability formula for Japanese texts by using a linear regression analysis which included indices such as the proportion of *hiragana* characters in the text, the mean number of predicates per sentence, the mean number of characters per sentence, and the mean number of *bunsetsu* boundaries² per sentence. As to question 2, much previous research thus far has adopted statistical methods, such as principal component analysis and regression analysis, applying them to Japanese text data that were formatted in specific ways.

The research presented in this paper aims at advancing text readability studies for the Japanese language and devises a practical and useful system that contributes to Japanese language teaching, learning, and research. More specifically, utilizing levelled corpora, mainly consisting of texts from Japanese textbooks³, we produced the following formula to measure the readability level of a given text in a six-level scale: $X = \{\text{mean length of sentence} * -0.056\} + \{\text{proportion of } kango * -0.126\} + \{\text{proportion of } wago * -0.042\} + \{\text{proportion of number of verbs among all words} * -0.145\} + \{\text{proportion of the number of auxiliary verbs} * -0.044\} + 11.724$ ($R^2=0.896$). The formula was tested against another set of levelled texts in Japanese to prove its reliability⁴. Lastly, the method was implemented in a computer system that calculates and produces the estimated level of a text via a web-based online interface.

It should be noted that the project presented in this paper is original in several ways. Firstly, the readability formula we constructed is intended especially for learners of Japanese as a foreign language, whereas many existing formulas such as those by Shibasaki and Hara (2010) and Sato (2011) are intended for native readers of Japanese. Secondly, our online implementation offers new functionalities that are not available in existing systems for reading support. These points are explicated in the following sections.

2 A *bunsetsu* is a unit of text in Japanese that is comprised of a content word plus the optional function word(s) that immediately follow it (Zhang and Ozeki 1998).

3 In the present paper, “Japanese textbooks” refer to “textbooks used for teaching Japanese to non-native learners”.

4 A *kango* is a Japanese word of Chinese-origin and thus is typically written in *kanji* characters, whereas a *wago* is a Japanese word that is neither loaned nor derived from words in a foreign language. A *wago* is typically written in *hiragana* or *kanji* characters in contemporary Japanese.

2. Data and methods

2.1 Overview

Two different sets of data were prepared for our research: model data and test data. The former consists of two types of text: one comprised of text from 83 Japanese textbooks, ranging from introductory to advanced, and the other comprised of text from National Diet meeting transcripts, selected according to the criteria explained in 2.2. From this basic data, we created corpora of six different levels. The readability measurement formula was produced by analyzing these levelled corpora. The latter dataset, that for testing the formula, consists of texts derived from 25 years of the Japanese-Language Proficiency Test (JLPT).

The levelled corpora for analysis were created from the original data in the following way. First, all texts were split into separate files of roughly the same size (around 1,000 characters). Second, each file was manually examined and then analyzed computationally and this enabled us to obtain corpora of six different levels. Then, each component text file in each of these levelled corpora was analyzed further using natural-language processing (NLP) tools, and various text features such as the frequency of words of different categories and different parts-of-speech were obtained. Using such numerical data as input values, a multiple linear regression analysis was conducted and, as a result, our readability measurement formula was finally obtained. The formula was then tested against a second dataset derived from JLPT, and its effectiveness was verified. The whole process is schematically summarized in Figure 1.

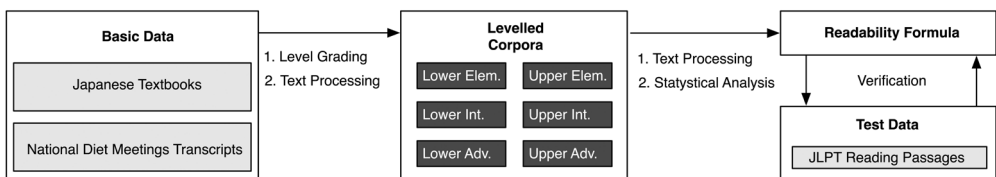


Figure 1. Data and procedures

2.2 Creating levelled corpora

The six-level scale we utilized throughout our research corresponds to lower-elementary, upper-elementary, lower-intermediate, upper-intermediate, lower-advanced, and upper-advanced levels. The model corpora of the first five levels were created using texts in Japanese textbooks, and that of the most advanced level was created from the text of National Diet meeting transcripts, which were included in the *Balanced Corpus*

of *Contemporary Written Japanese* (BCCWJ)⁵. The total number of words in the levelled corpora is 595,360. Table 1 shows how these texts were divided⁶.

Table 1. Basic statistics of the levelled corpora

	Lower- elem. (133)	Upper- elem. (117)	Lower- int. (148)	Upper- int. (286)	Lower- adv. (117)	Upper- adv. (194)
Word types	3,178	2,858	5,156	10,291	6,833	4,712
Word tokens	72,691	68,746	87,433	174,953	69,268	122,269

*Numbers inside parentheses represent the number of text passages included

The actual procedure for grouping the original data into these levels was comprised of three steps. First, the first author of the present paper checked the general design (such as purpose, contents, and featured study items) of each of the textbooks in the original dataset, and categorized them into five levels from lower-elementary to lower-advanced. Second, we asked three practicing teachers of Japanese to manually examine the text passages thus categorized and choose only those that they thought truly matched the given level. Finally, the results were further verified using the statistical method of discriminant analysis.

2.2.1 Choice of data and data size

There are two supplementary comments on the basic statistics of the levelled corpora presented in Table 1. The first concerns the choice of the original data, and the second concerns data size.

The decision to use Japanese textbooks to construct a corpus for each of the five levels from lower-elementary to lower-advanced was motivated by the following: in text readability studies, it is required that a clear indication of the level of the model data be already given so that a formula can be drawn by analyzing it. Thus, it has traditionally been the case that readability research uses language textbooks. The reasoning behind this is obvious: textbooks are written according to the assumed levels of the readers who use them. The vocabulary, idioms, structures, and types of logic used in textbooks of

5 http://pj.ninjal.ac.jp/corpus_center/bccwj/en/

6 One may expect vocabulary variation (number of word types) to increase as the level of difficulty increases. In Table 1, however, there are less word types in the upper-advanced corpus (Diet transcripts) than in the lower-advanced (textbooks), even though it is larger than the lower-advanced corpus. This is probably due to the fact that Diet transcripts repeatedly deal with a rather limited set of topics. Another possibility is that sentences in the Diet transcripts tend to be composed by combining two or more clauses with conjunctions, and as a result they contain a relatively larger number of functional words than sentences of other types of text.

different grades are, in general, fairly controlled. We find this characteristic of language textbooks ideal for our purpose. In fact, however, there are some researchers who see language in textbooks as unnatural, or at least somehow different from language observed elsewhere. This is actually a matter of degree and the same can be said about written language of any kind. We concluded that the benefit of using textbooks exceeded any possible drawbacks.

We used National Diet meeting transcripts for our highest-level corpus based on the following four motivations. First, these are transcripts of genuine utterances, and are not artificially created data. This results in a variety of styles in the data, which is often considered characteristic of a highly advanced set of linguistic data. Second, this approach provided a sufficient amount of text. As shown in Table 1, the number of words for this level is comparable to those of other levels, even if not necessarily exceeding them. Third, the sentences used are relatively long, which is broadly considered a condition for an advanced text. Finally, the fourth reason was that the data contained utterances dealing with abstract concepts and ideas. For these reasons, and also taking into consideration the facts observed in texts of different registers by Lee (2011), we made a decision to exclusively use National Diet meeting transcripts to compile the corpus associated with our upper-advanced level. Lee (2011) carried out a close examination of the text in National Diet meeting transcripts and showed that it should be placed well beyond the level of texts used for JLPT L1 tests (highest-level).

Another point that requires comment is that the data size of the five corpora from lower-elementary to lower-advanced is not balanced, as is apparent in Table 1. This is due to the fact, firstly, that there is a relatively larger number of available titles of textbooks at the intermediate level. Secondly, elementary level textbooks contain shorter sentences and they accordingly have fewer words. A third reason is that there are only a limited number of available titles for advanced learners. Thus, the data size of the corpora at different levels is different. Still, each corpus has a fairly large amount of text, and the effect of size difference among corpora was considered very small, if any.

2.2.2. *The rationale behind these six levels*

So far we have not presented a sufficient explanation as to why all texts were split into files of approximately 1,000 characters and why we adopted the six level-scale in the first place.

The reason behind creating text files of roughly the same size had much to do with the fact that in text readability studies, various indices regarding “length” observed in a text have much to do with the level of the text. Such indices include the mean length of sentences, the mean length of words, and the total number of words in a text. It is essential for text readability research to make certain that such indices are retrieved as

accurately and efficiently as possible.⁷ Thus, standardizing text size is a prerequisite for obtaining characteristics regarding readability. The choice of 1,000 for the number of words contained in one text file is, however, rather arbitrary. It is not necessarily based on a specific scientific fact. Rather, it is motivated by the fact that in many Japanese language courses for non-native speakers a text of about 1,000 characters is typically preferred because this fits in well with the duration of a lesson.

We categorized the model text files into six levels. It would have been possible for us to choose two or three category levels instead, as many textbooks are simply levelled as “elementary,” “intermediate,” and “advanced,” but did not for several reasons. Firstly, a textbook teaching Japanese often contains materials of different levels; the level of the very first chapter in the book can be largely different from that of the last chapter in the same book. Actually, this is quite a natural phenomenon, as textbooks are designed so that their users’ ability gradually increases as the pages proceed. This fact urged us to break up one single textbook into parts and put them into different categories according to the specific content level. Secondly, the common practice of dividing textbooks into “elementary,” “intermediate,” and “advanced” is not necessarily rigidly standardized among publishers and authors. Some textbooks adopt a level system based on the frequency of the use of complex grammatical constructions, but others adopt a level system based exclusively on the vocabulary items used. For these reasons, we devised a six-level scale, which does not exclusively depend upon either grammatical or lexical characteristics, nor did we risk placing texts into two or three haphazard levels.

As a result of splitting the dataset into files of about 1,000 characters, we obtained 995 text files in Japanese. Levelling them was not, however, necessarily completed by reading text files and manually sifting them one-by-one. This would be not only time and resource consuming, but also highly ineffectual. Thus, we devised a method of text categorization that made use of both human graders and computational tools. First, the textbook dataset was roughly sorted into five levels from lower-elementary to lower-advanced by one of the authors of this paper (mainly according to the general facts already known about the titles). Then we asked three teachers of Japanese who have more than 10 years of teaching experience to examine all the files in each of the levelled file pools that were created in the pre-categorization process. They were then asked to pick out exactly 30 files that they thought contained texts quite representative for each of the five levels. Then we selected 20 files that were chosen by multiple graders for each level, creating a subset of the original dataset that was comprised of five groups of 20 files, each of which is thought to be more or less prototypical of each level. Furthermore, we

7 As is pointed out by some of the pioneers in this field such as Flesch (1948), Sakamoto (1964), and Smith and Kincaid (1970), the larger the length of a sequence, whether it is of a sentence or a word, the heavier the burden on working memory. The readability of a text is roughly in negative correlation with the mean size of various textual elements.

carried out a discriminant analysis (described in more detail in 3.1), using these core data as a model, against the text files that had been “filtered out” in the previous process, and finally obtained the levelled corpora of text, each of which contained not only 20 files, but also files that supposedly have similar textual characteristics to those of the core data. Table 2 shows descriptions of the assumed abilities of readers of each level given to the graders before they examined the texts.

Table 2. Descriptions of readers’ reading abilities for six levels

Level	Description
Upper-advanced	The reader is able to fully understand highly technical writing. S/he has no difficulty dealing with virtually any kind of text in Japanese.
Lower-advanced	The reader is able to mostly understand technical writing. S/he can deal with complex structures often observed in literary works.
Upper-intermediate	The reader is able to grasp the overall structure of technical writing. S/he can deal with Japanese texts found in most day-to-day situations without much difficulty.
Lower-intermediate	The reader is able to read relatively simple writing and can deal with texts comprising multiple sentences.
Upper-elementary	The reader can understand basic vocabulary items and grammatical patterns. S/he can deal with complex sentences of basic types such as ones involving <i>-te</i> form.
Lower-elementary	The reader can understand the most fundamental Japanese expressions used in simple sentences. S/he has difficulty in dealing with complex sentences or sentences containing adnominal modifiers.

The following passages are samples of the core data collected as a result of the process mentioned above for the five category levels from lower-elementary to lower-advanced, and a sample of text of the upper-advanced level, which is from National Diet meeting transcripts.

1) Lower-elementary

音楽が好きですから、よくCDを聞きます。日本が好きですから、日本語を勉強します。安かったですから、買いました。ディズニーランドは楽しかったです。教室は静かでした。わたしはラーメンが好きです。わたしはたばこがきらいです。ワンさんは日本語が上手です。わたしは料理が下手です。

I like music, so I often listen to CDs. / I like Japan, so I study Japanese. / I bought it because it was inexpensive. / I had fun at Disneyland. / It was quiet in the classroom. / I like ramen. / I hate cigarette smoke. / Mr. Wang is good at speaking Japanese. / I am not good at cooking.

2) Upper-elementary

わたしは夏休みに国へ帰らないつもりです。わたしは30歳まで結婚しないつもりです。わたしは大学へ行かないつもりです。わたしは学校では母国語を使わないつもりです。わたしは車に乗らないつもりです。今年の夏も国へ帰りますか。はい、そのつもりです。いいえ、帰らないつもりです。

I am not going to my home country during the summer holiday. / I will not marry until I am thirty.

I am not going to a university. / I will not use my native language at school. / I will not drive a car. / Are you going to your country in the summer again this year? Yes, I am. / No, I'm not.

3) Lower-intermediate

毎週1回は祖母の家に子どもたちが孫たちをつれて集まります。とてもにぎやかです。祖母の80さいの誕生日には、マニラで一番大きなホテルを借りて、大家族の全員と親しい友人が、全部で500人以上集まりました。ごちそうを食べたり、ダンスをしたり、歌をうたったりして、とてもにぎやかでした。祖母もワルツやチャチャチャをおどりました。それから子どもと孫の全員が花をプレゼントしました。

My grandmother has all her children and grandchildren gather at her house once a week. Her house is filled with laughter and lively conversations. On her 80th birthday, we held a big party at the biggest hotel in Manila with more than 500 participants, including all of her family members and friends. At the boisterous party, we enjoyed wonderful food, sang songs together, and danced to the music. Grandma herself danced a waltz and the cha-cha. At the close of the party, all her children and grandchildren presented her with flowers.

4) Upper-intermediate

今でいうリフォーム、リサイクルをごく当たり前のこととしてやっていました。日本は、1950年代後半から経済の成長がいちじるしく、供給がどんどん増加し、国民一人あたりの所得も上がってきました。この時代を境にして、需要と供給のバランスが逆転しました。現在の日本は完全に供給が過剰、需要が不足している時代です。ものをつくる企業はこういうときにどうするでしょうか。

Back then, many Japanese people were already doing what we now call “reforms” and “recycles.” However, the economy grew so rapidly in the 1950s that there was a tremendous increase of supply, and the average income earned per person rose markedly. It was the time when the supply-demand unbalance gradually set in. Now, Japan is in a state of excess supply with limited demand. We should think about what manufacturers can do under such circumstances.

5) Lower-advanced

「実現への戦略—それは産業界を納得させる手順」—土地税制改革のため、大蔵省が省内論議をまとめた内部資料に、こんなくだりがある。「土地などの資産所得が勤労所得よりはるかに大きくなり、勤労意欲を低下させて日本の経済・社会基盤を揺るがしている」「いや、資産効果で消費は拡大、重厚長大産業は土地を活用して企業基盤を強くしている面もある」...連日の、こんな議論を経て土地保有税創設に動き出した大蔵省が、土地税制改革のポイントは対財界戦略にあるとみていることを示す文言だった。

“The Strategy for Realization: A Procedure to Persuade Business Sectors”—this is a phrase found in one of the Finance Ministry’s internal documents that contain summaries of discussions on the land-tax system. “For many people, the income from their real estate assets has become far larger than their earnings from work, and the decrease in their incentives to work is starting to largely influence the economic and social foundations of Japan.” “It does not necessarily explain the whole picture—the asset effect has increased consumption, and the heavy industries are strengthening the corporate infrastructure by making good use of the land they own.” The Finance Ministry, after repeated discussions like this, is currently doing preparatory work for creating a land-holding tax. They consider it extremely important to be highly strategic in negotiating with the major players in the business community, as is apparently shown in the above-mentioned passage.

6) Upper-advanced

「あの際の米軍による行動が、イラクに関連する一連の国連安保理決議の履行を確保するため、それに必要な措置ということであれば、我が国としてはこれを理解し、支持する、こういうことを申したわけでございます、我が国としてはいわば無条件で米国のやることはすべて支持しますよということとは申し上げておりません。御承知のとおり、国連には一連の決議がございます、イラク軍が北部イラク地域から撤退するよにということはずっと国連として求めておったわけでございます。そういったことが確保されるために必要な措置ということ米軍が行動するのであれば、それは理解し、支持する、こういうことを明らかにしたということでございます。」

“What I intended to state was only that our country should understand and support the conduct of the American Armed Forces on that occasion because it was a necessary procedure to make sure the series of UN Security Council Resolutions on Iraq were implemented; I was not saying that our country would endorse everything that the US does without any conditions. As you know very well, the UN has made a series of resolutions, and they clearly requested that Iraqi forces withdraw from the Northern Iraq regions. If the US Armed Forces were conducting a necessary action to ensure the implementation of the UN resolutions, then the Japanese government should understand and support it. That is what I intended to make clear then.”

2.3 Selection of formula variables

In order to construct a formula to calculate the readability of Japanese texts, firstly it was necessary to analyze our model data with NLP tools. Thus, we analyzed our dataset using the Japanese morphological analyzer MeCab 0.996 with UniDic 2.2.0.⁸ Obtained from this process were types of data such as: 1) mean length of sentence, 2) proportion of nouns, 3) proportion of auxiliary verbs, 4) proportion of verbs, 5) proportion of subsidiary verbs, 6) proportion of adjectives, 7) proportion of *wago* words, and 8) proportion of *kango* words. We selected these elements based on work by Shibasaki and Hara (2010), as candidates for variables to be used in our formula.

In our selection of elements for use as variables, there were limitations that needed to be considered. Firstly, since the resulting formula would be computationally implemented in a web-based readability measurement system, only values that could be immediately calculated were available to us. In reality, there could be numerous variables that affect the readability of texts. Theoretically, it is conceivable that there are not only

⁸ MeCab (<http://taku910.github.io/mecab/>) can be used with one of several available dictionary packages, of which UniDic is one option (<http://osdn.jp/projects/unidic/>). UniDic is superior to other dictionaries in that the format of its entry items is systematically standardized based on short-unit words (SUW) and it offers richer lexical information including that of word types regarding etymological origins (*wago*, words of Japanese-origin; *kango*, words of Chinese-origin; or *gainaigo*, words of Western-origin). See Den (2009) for further details about UniDic.

purely numerical ones such as the frequency of certain type of words, but also those that represent more abstract aspects of texts such as the overall cohesion, the stylistic tone of the text, or even the size of font type and the color of a printed text. However, we had to exclude from our formula those types of information that are difficult to obtain computationally, even though some might be effective in determining the real readability of a text.

Secondly, although using an NLP dependency analysis tool could be helpful for producing an accurate formula, it was not a realistic option. In fact, Shibasaki and Hara (2010) used the results of dependency parsing in their model. Tools for dependency parsing are currently available, including ones that were adopted by Shibasaki and Hara (2010)⁹. However, they suffer from a problem of insufficient accuracy (more than 10 percent of text is analyzed incorrectly). Thus, we decided not to use this type of technology in constructing our formula and the web-based system we built based on the formula.

Thirdly, we chose to use only variables that are proportional, instead of those that are numerically absolute. The output of a formula that adopts the latter types of variables would be significantly influenced by the size of the input text. This makes it difficult to compare readability scores for texts of different sizes. By using only proportional frequencies, we can measure the readability of texts of any size and we can make sure that the resulting scores are comparable to each other.

The formula was constructed with linear regression analysis. Linear regression analysis is a statistical method that has also been used in past readability studies (e.g., Tateishi et al. 1988; Shibasaki and Hara 2010). It is helpful when explaining the correlation among two or more variables based on a linear model. We conducted multiple linear regression analysis using IBM SPSS (ver. 22).

2.4 About test data

In addition to the levelled corpora based on the core dataset described above, we also built a test corpus comprised of text files other than those contained in the latter to confirm the validity and the reliability of the formula.

There is an important fact to note regarding the test data. The levels estimated for input texts using our formula do not necessarily have pre-existing external criteria. In fact, this is the case with virtually every attempt in text readability measurement. Suppose, for instance, one desires to measure the readability of a Japanese newspaper article by applying a readability formula to the text and obtains an estimated upper-advanced level. How do we verify that the result is correct, or reject it as incorrect? As such, readability levels are to some extent inevitably subjective. Thus, the verification of the readability formula is not necessarily an easy task.

⁹ Shibasaki and Hara (2010) used CaboCha, a Japanese dependency structure analyzer (<http://taku910.github.io/cabocha/>).

To minimize such concerns and also to verify that the application of our formula was as reliable and usable as possible, we constructed test corpora using texts from reading passages in JLPT from 1984 to 2008. The division of the data is presented in Table 3.

Table 3. Test corpora

Level	Number of words	Mean number of words per sentence
L1 (78)	50,511	28.3
L2 (66)	42,586	24.5
L3 (17)	10,541	16.4
L4 (11)	6,242	10.9

* Numbers inside parentheses represent the number of text passages included

As in the case of the model data, the test data consisted of text files, each of which contained around 1,000 characters. The L1 level (highest-level) corpus had 50,511 words in total and was comprised of 78 files. The corpora of other levels were constructed in the same fashion. Also, as in the case of the model data, the higher the level, the greater the number of words in the corpus. This is mainly because the JLPT tests for more advanced levels have longer sentences than those for lower levels. This is apparent from the mean number of words per sentence in each of the test corpora: 28.3 for L1, 24.5 for L2, 16.4 for L3, and 10.9 for L4.

The test was carried out by examining the degree of match between the test corpora and the estimated levels obtained by applying the data in the test corpora to our formula.

3 Results and discussion

This section describes the procedures and results of the analysis in further detail. 3.1 presents a closer look at the way the levelled data of the model corpora were constructed, explaining how the division of the corpora was drawn from the discriminant analysis. In 3.2, the results of the multiple linear regression analysis carried out to construct the formula are expounded. And in 3.3, the results of the verification of the formula using the test data are presented.

3.1 Results of the discriminant analysis: Constructing the levelled corpora

As briefly described previously, we manually classified the original data and then extracted 20 text files containing data that assumedly matched each of the six levels from

lower-elementary to upper-advanced. The resulting “core” data of 120 files were utilized to classify the other 875 files, that is, the rest of the original dataset of 995 text files, using discriminant analysis. As a result, for the lower-elementary level, 78 text files were re-selected out of 113 files that had been rejected from the core data by graders by way of manual examination. Similarly, 37 files out of 97 files for upper-elementary, 58 files out of 128 files for lower-intermediate, 102 files out of 266 files for upper-intermediate, 60 files out of 97 files for lower-advanced, and 152 files out of 174 files for upper-advanced that had been once rejected by graders were re-selected for the respective levels as presented in Table 4.

Table 4. Discriminant analysis results

		Levels predicted by discriminant analysis						Total
		Upper-adv.	Lower-adv.	Upper-int.	Lower-int.	Upper-elem.	Lower-elem.	
Original levels	Upper-adv.	152	14	8	0	0	0	174
	Lower-adv.	6	60	24	7	0	0	97
	Upper-int.	8	70	102	61	22	3	266
	Lower-int.	0	4	39	58	21	6	128
	Upper-elem.	0	1	14	28	37	17	97
	Lower-elem.	0	0	0	7	28	78	113
Total		166	149	187	161	108	104	875

Finally, among the 995 text files contained in the original dataset, 607 were used to construct the levelled corpora and the other 388 files were filtered out, as the latter files were not grouped to levels either in the selection process by human graders or the discriminant analysis.

3.2 The readability formula

The readability formula was selected from five models generated as a result of multiple linear regression analysis. Figures involved in the analysis are shown in Table 5.

Table 5. Multiple linear regression analysis results

Models		Coefficient	R ²
Model 1	(Constant)	5.938	0.787
	Mean length of sentence	-0.099	
Model 2	(Constant)	6.691	0.839
	Mean length of sentence	-0.082	
	Proportion of <i>kango</i>	-0.073	
Model 3	(Constant)	13.195	0.878
	Mean length of sentence	-0.063	
	Proportion of <i>kango</i>	-0.153	
	Proportion of <i>wago</i>	-0.086	
Model 4	(Constant)	12.128	0.893
	Mean length of sentence	-0.057	
	Proportion of <i>kango</i>	-0.142	
	Proportion of <i>wago</i>	-0.061	
	Proportion of verbs	-0.159	
Model 5	(Constant)	11.724	0.896
	Mean length of sentence	-0.056	
	Proportion of <i>kango</i>	-0.126	
	Proportion of <i>wago</i>	-0.042	
	Proportion of verbs	-0.145	
	Proportion of auxiliary verbs	-0.044	

Among the five models constructed by using the multiple linear analysis in Table 5, Model 1 is the simplest. It is composed only of a constant and the mean length of sentences. Its R², an index that shows prediction accuracy, is 0.787. Model 2 includes the proportion of *kango*, words of Chinese origin, in addition to a constant and the mean length of sentences, with its R² being 0.839. Having examined Models 3 to 5 in the same token, the R², the coefficient of determination, of each of the 5 models is plotted as in Figure 2.

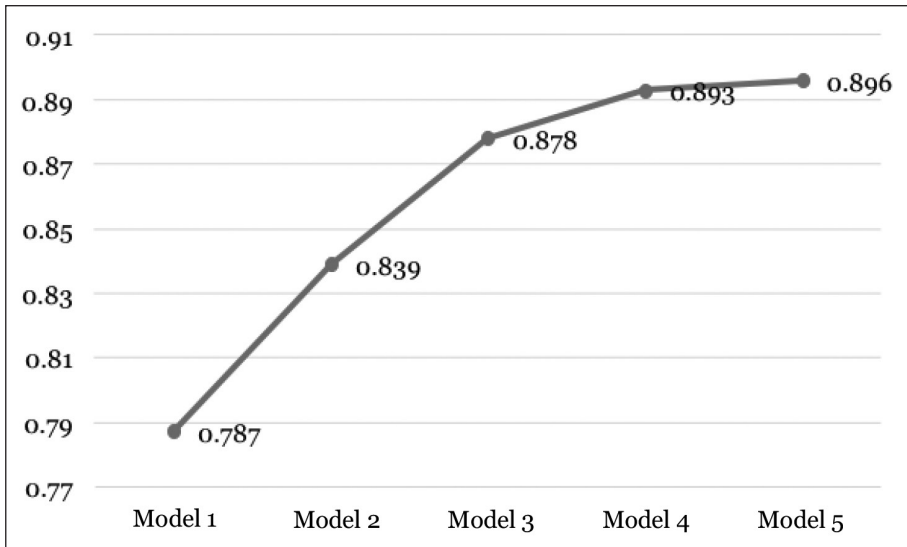


Figure 2. Transition of the coefficient of determination

Among the five models, Model 5 was finally selected as it showed the highest prediction accuracy. Based on this model, the following readability formula was obtained.

Readability Formula for Japanese Language Education ($R^2 = 0.896$)

$$X = \{\text{mean length of sentence} * -0.056\} + \{\text{proportion of } kango \text{ words} * -0.126\} + \{\text{proportion of } wago \text{ words} * -0.042\} + \{\text{proportion of number of verbs among all words} * -0.145\} + \{\text{proportion of number of auxiliary verbs} * -0.044\} + 11.724$$

The formula shows that three indices are especially effective when measuring the readability level of a Japanese text (for language education). First among them is the mean length of sentence, as would be naturally expected. It is considered that this indirectly reflects the degree of structural complexity of a sentence in the text passage. Secondly, the proportions of *kango* and *wago* are effective. This is considered to be due to the fact that many words of technical and/or abstract concepts tend to be realized as *kango*, whereas most *wago* are considered more basic and fundamental. And thirdly, the proportion of verbs and the proportion of auxiliary verbs are also effective. It is assumed that these two indices reflect, again, the degree of structural complexity of the text. For a more concrete example, our formula is applied to a sample text of the lower-elementary level presented in 2.2 as follows:

$$\{8.56 * -0.056\} + \{9.09 * -0.126\} + \{63.64 * -0.042\} + \{2.60 * -0.145\} + \{22.08 * -0.044\} + 11.724 = 6.08$$

The resulting score, 6.08, can be interpreted using a correspondence table as in Table 6. It is within the range of 5.5 to 6.4, thus the text is interpreted as lower-elementary.

Table 6. Levels and readability scores

Level	Readability score range
Upper-advanced	0.5 - 1.4
Lower-advanced	1.5 - 2.4
Upper-intermediate	2.5 - 3.4
Lower-intermediate	3.5 - 4.4
Upper-elementary	4.5 - 5.4
Lower-elementary	5.5 - 6.4

There is a caveat. The resulting readability score could be smaller than 0.5, the lower limit on the table, or larger than 6.4, the higher limit. When such a case arises, then the text can be considered to have some characteristics that our formula cannot properly deal with. For example, an extremely short text that includes many *kango* in long sentences could produce a score less than 0.5. On the contrary, a text passage having many *wago* in extremely short sentences could produce a score over 6.4. In any case, such instances are rightfully considered exceptional when dealing with texts for Japanese reading education.

3.3. Verification results using test data

In this section, the results of verification using the test data introduced in 2.4 are presented. The logic behind the procedure is this: if readability scores produced by applying the formula to texts from JLPT tests, which have already been levelled, predict the text levels sufficiently correctly, then the formula is considered highly valid. The resulting figures of this experiment are summarized in Table 7.

Table 7 presents a cross tabulation of JLPT levels of the test data, on the one hand, and the estimated readability levels calculated using the formula, on the other. Several things can be noted here: 1) the reading passages in JLPT L1 are mostly estimated to be of upper-intermediate or lower-advanced, 2) the reading passages in JLPT L2 are mostly estimated to be lower-intermediate or upper-intermediate, 3) the reading passages in JLPT L3 are exclusively estimated to be upper-elementary or lower-intermediate, and 4) the reading passages in JLPT L4 are exclusively estimated to be lower-elementary or upper-elementary.

Now let us examine the results of the same experiment in the form of numeral scores, instead of discrete levels. Figure 3 represents the distribution of the scores in the form of a

Table 7. Cross tabulation of JLPT levels and levels estimated using the formula

			Estimated readability level					Total
			Lower- elem.	Upper- elem.	Lower- int.	Upper- int.	Lower- adv.	
JLPT Level	L1	Num. of passages	0	0	6	47	25	78
		%	0.0	0.0	7.7	60.3	32.1	100.0
	L2	Num. of passages	0	1	19	44	2	66
		%	0.0	1.5	28.8	66.7	3.0	100.0
	L3	Num. of passages	0	7	10	0	0	17
		%	0.0	41.2	58.8	0.0	0.0	100.0
	L4	Num. of passages	5	6	0	0	0	11
		%	45.5	54.5	0.0	0.0	0.0	100.0
Total	Num. of passages	5	14	35	91	27	172	
	%	2.9	8.1	20.3	52.9	15.7	100.0	

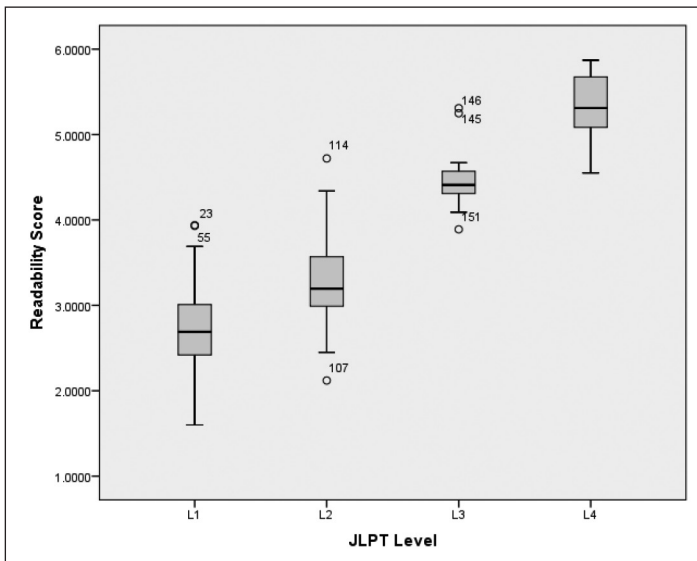


Figure 3. Estimated readability levels of text in JLPT L1 to L4

box plot. The figure shows that the larger the JLPT level-number, the higher the readability score estimated by our formula (Note that a larger JLPT level-number represents a less advanced test level, and a higher readability score means the text in question is relatively easy). One-way analysis of variance showed that the difference among the four groups in terms of their mean numbers is statistically significant ($F(3, 168) = 141.035, p < 0.001$).

Another important fact noted for Figure 3 is about the overall tendency of the results. According to the estimated levels worked out by the calculation using our formula, while L1 and L2 show a relatively small gap between them, the gap between L2 and L3 is larger. It is also larger than the gap between L3 and L4. In fact, this conspicuous gap between text passages of L2 and those of L3 has been known among people involved in the test and has been addressed in the new version of JLPT that is divided into 5 levels. The present experiment finally attests to this.

In concluding this section, the estimated scores of text (and accordingly the levels) obtained using our formula with the JLPT reading passages largely correspond to the original JLPT text levels. This confirms the high reliability of the formula gained as a result of the present research.

4 Web system implementation

4.1 Overview

As an attempt to utilize the output of our research presented thus far, we developed a web system that estimates readability scores and levels of texts in Japanese. The system is currently available at <http://jreadability.net>. We expect that primary users will be practicing teachers of Japanese who need to prepare reading materials for classes to match student levels. Our system also makes several features available that will be helpful not only to teachers, but also to learners.

There are existing systems available for automatic readability assessment such as those developed and introduced in Shibasaki and Hara (2010) and Sato (2011), but they are built on corpora of textbooks written in Japanese for native speakers of Japanese; their formulae consequently assess the readability of Japanese texts on a scale corresponding to Japanese school grades, and as such are not directly applicable to selecting texts for readers of Japanese as a foreign language. On the other hand, our formula is built on levelled corpora of textbooks for learners of Japanese as a foreign language. Therefore, it is expected to be easier to use for both teachers and learners of Japanese.

4.2 Basic system design

In order to calculate readability scores and levels from input texts using our formula, the system needs to first parse the text into words. The input text is split into sentences

by a full-stop symbol and then each of these sentences is further split into words. Since word boundaries in Japanese texts are not indicated by spaces, splitting sentences into words requires an NLP tool called a morphological analyzer. To create a system that is capable of accomplishing this in the same fashion as when we dealt with corpus data to extract lexical information in the process presented in 2.3, we adopted the same set of equipment, MeCab (0.996) and UniDic (2.2.0). With these tools working on the backend, the system extracts five numerical indices from the input text: 1) mean length of sentences, 2) proportion of *kango*, 3) proportion of *wago*, 4) proportion of verbs, and 5) proportion of auxiliary verbs. The system applies these values to our formula to obtain the readability score.

Figure 4. Input form of online readability measurement system

Figure 4 is a screenshot of the text input form of this online system. The user inputs the text and presses the 実行 *jikkō* ('run') button. The results are immediately presented as shown in Figure 5.

As mentioned above, although the calculation of the readability score needs values for only five types of variables, other types of data obtained as a result of the text analysis using MeCab and UniDic are also presented. Among those are the total token number of words and the total type number of words of the input text, as well as the frequency and distribution of vocabulary items of different levels, the frequency and distribution of vocabulary items of different parts-of-speech, and the frequency and distribution of vocabulary items of different types of origin (such as *wago*, *kango*, and *gairaigo*) as shown in Figure 5.

テキストの概要

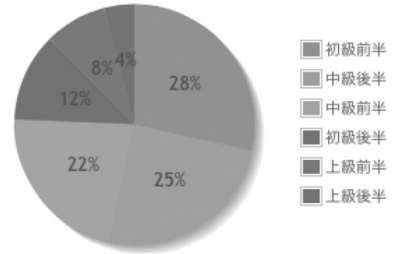
総形態素数（異なり）を表示するには「語彙リストを出力」をオンに

文章難易度	上級前半
リーダビリティ・スコア	2.15
総文数	12
総形態素数（延べ）	458
総形態素数（異なり）	184
総文字数（記号・空白を含む）	741
一文の平均語数	38.17

語彙レベル構成

語彙レベル情報を持っている形態素だけを集計

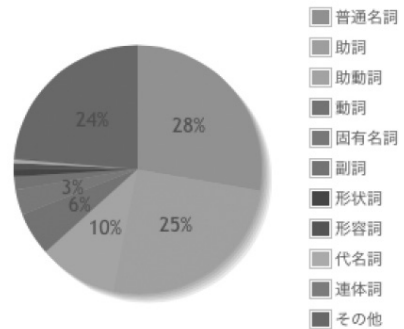
初級前半	55
中級後半	48
中級前半	43
初級後半	23
上級前半	16
上級後半	8



品詞構成

記号類は除外

普通名詞	127
助詞	116
助動詞	47
動詞	26
固有名詞	15
副詞	8
形状詞	4
形容詞	4
代名詞	2
連体詞	1
その他	108



語種構成

定型句は「ありがとう」などを指す

和語	276
漢語	108
外来語	10
混種語	4
定型句	0

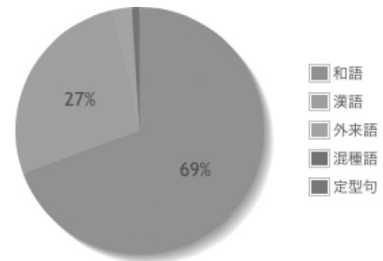


Figure 5. Sample results of readability measurement

4.3 Additional features

The statistics and graphs in Figure 5 are presented on the pane with a tab titled テキスト情報 *tekisuto jōhō* ('Text Information'). There are two other tabs next to it, one of them being テキスト詳細 *tekisuto shōsai* ('Text Details'), and the other 語彙リスト *goi risuto* ('Vocabulary List'). Selecting テキスト詳細 *tekisuto shōsai*, the user is presented with the input text, with its component sentences sequentially numbered and words highlighted with different colors according to the vocabulary level as shown in Figure 6.

The screenshot shows a web interface with four tabs: '本システムについて', 'テキスト情報', 'テキスト詳細', and '語彙リスト'. The 'テキスト詳細' tab is active. Below the tabs, there are buttons for '結果保存 (CSV: Shift-JIS)' and '結果保存 (CSV: UTF-8)'. The main content area displays 12 numbered sentences. Each sentence contains words highlighted in different colors, corresponding to six vocabulary levels: 初級前半 (light blue), 初級後半 (light green), 中級前半 (light orange), 中級後半 (light red), 上級前半 (light purple), and 上級後半 (light pink). The text is in Japanese and discusses environmental and breeding topics.

Figure 6. Text details

The system has a levelled vocabulary list for learners of Japanese in its background that was produced by Sunakawa et al. (2012). The list consists of six sub-lists of different levels (lower-elementary, upper-elementary, lower-intermediate, upper-intermediate, lower-advanced, upper-advanced).

A similar feature is already available in the reading support system Reading Tutor (Kawamura 1999)¹⁰. However, while Reading Tutor categorizes vocabulary according to

10 http://language.tiu.ac.jp/index_e.html

the 4 levels of the old version of JLPT, our system uses a more fine-grained six-level vocabulary list, which is expected to be more easily applicable to actual learning environments. Moreover, the system also includes a built-in dictionary with definitions and example sentences. Inside our system, each of the words in the input text is checked to see if it is included in one of the sub-lists of the levelled vocabulary list. If this is the case, the word is highlighted with a color according to the level. When one of the highlighted words is clicked, a pop-up window will appear showing dictionary definitions and example sentences of the word, which were also provided as a product of Sunakawa et al. (2012).

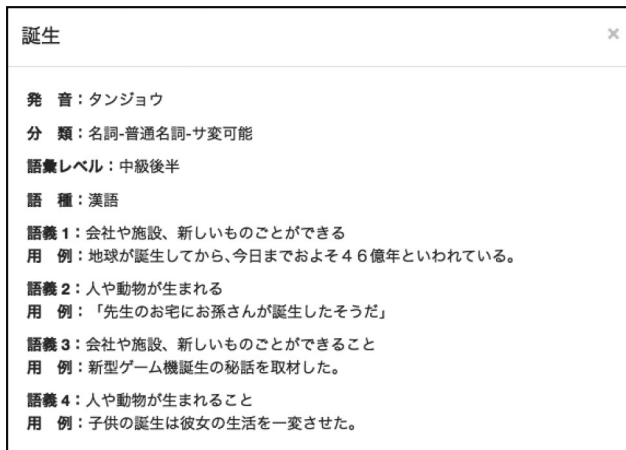


Figure 7. Pop-up window showing definitions and examples

The features presented above will be helpful for teachers of Japanese and also for learners. Other features for learners implemented in this system include the text read aloud with synthesized voice. Once a readability measurement process has been completed, a headphone-like little icon appears above the text input form if the web-browser being used is natively capable of the text read aloud. Clicking on this icon will play the input read aloud text in a synthesized voice.¹¹

Our web system has some other features that may also be helpful to researchers of Japanese as a foreign language, or Japanese linguistics in a broad sense. Once a readability measurement process has been completed, a 語彙リスト *goi risuto* ('Vocabulary List') tab appears. Clicking on the tab, the user will be presented with a list of all the words in the input text as in Figure 8. The data are aggregated into their basic forms (e.g., 取り組む *torikumu* ('work on') is the basic form for variations such as 取り組み *torikumi* or

11 As of writing, not many web-browsers support the Web Speech API, which our system depends on for its read aloud functionality. Currently, we have only tested this functionality on Google Chrome, one of a few browsers that support the API.

取り組ん *torikun*) and include the following types of related information, by which the user can sort and rearrange the data on their web-browser. These data are downloadable in the comma-separated value (CSV) format.

Basic form	取り組む	(<i>torikumu</i>)
Pronunciation	トリクム	(<i>torikumu</i>)
Grammatical category	動詞-一般	(<i>verb-general</i>)
Surface form(s)	取り組ん	(<i>torikun-</i>)
Frequency (%)	2 (0.44%)	
Vocabulary level	中級後半	(<i>upper-intermediate</i>)

出現順	基本形	読み	分類	出現順で並べ替え	読みで並べ替え	分類で並べ替え	頻度で並べ替え	語彙レベルで並べ替え
1	環境	カンキョウ	名詞-普通名詞-一般	3	0.66	環境 (3)		中級後半
2	省	ショウ	接尾辞-名詞的-一般	3	0.66	省 (3)		中級後半
3	が	ガ	助詞-格助詞	16	3.49	が (16)		
4	初めて	ハジメテ	副詞	4	0.87	初めて (4)		初級後半
5	取り組む	トリクム	動詞-一般	2	0.44	取り組ん (2)		中級後半
6	で	テ	助詞-接続助詞	2	0.44	で (2)		
7	いる	イル	動詞-非自立可能	3	0.66	いる (3)		初級前半
8	国	クニ	名詞-普通名詞-一般	2	0.44	国 (2)		初級前半
9	の	ノ	助詞-格助詞	27	5.9	の (27)		
10	特別	トクベツ	形状詞-一般	1	0.22	特別 (1)		中級前半
11	天然	テンネン	名詞-普通名詞-一般	1	0.22	天然 (1)		中級後半
12	記念	キネン	名詞-普通名詞-サ変可能	1	0.22	記念 (1)		中級後半
13	物	ブツ	接尾辞-名詞的-一般	1	0.22	物 (1)		中級後半
14	,		補助記号-読点	23	5.02	、 (23)		

Figure 8. Vocabulary list (partial)

4.4 System limitations

The online system allows a user to measure the readability of a Japanese text and also offers many functions useful to educators, learners, and researchers of the language. There are, however, some limitations that the user should note. Firstly, depending on the nature of the input text, the system may not perfectly parse the text and break it into individual words in the most appropriate way. The model data used to devise the readability formula are mostly from textbooks of Japanese. The NLP tools are able to analyze such text easily because it does not have many neologisms; it is mainly composed of words that are well established in the language. The online-system we developed,

however, is required to analyze whatever type of text that the user inputs. Accordingly, the text could be of various types such as a piece of text written especially for elementary learners using quite a limited number and variety of words, or a blog text containing many newly-coined words and/or highly technical terms, which would be difficult for the NLP tools to handle properly.

A second limitation concerns the morphological analysis completed using the NLP tools. Normally, texts in Japanese do not have intervening spaces to make the boundaries of words visible. Morphemes are combined with each other forming larger units, namely words. They are combined with each other with different strengths, making the distinction between morphemes and words less clear. Thus, there are a couple of different ways in which the size of a word-unit is determined for Japanese text. We adopted short-unit words (SUW) among other possible word-units such as long-unit words (LUW) mainly because of the specifications of the NLP tools we used. With SUW, a sequence such as 環境省 *kankyōshō* ('environment ministry') is analyzed as two individual words 環境 *kankyō* ('environment') and 省 *shō* ('ministry') sequentially arranged back-to-back. Some users might find it slightly unnatural since what is referred to by this sequence of two morphemes is just one single concept, or institution. They may prefer to have such a sequence treated as a single (compound) word, rather than as two individual elements.

The latter limitation is, however, mostly at a presentation level, and it does not significantly affect the readability measurement. It is possible that future enhancements and improvements of the NLP tools will enable us to repeat the same set of procedures as described in the present paper to devise a possibly better readability formula based on a different type of unit of words such as LUW. The current formula based on SUW nonetheless has been proven effective as presented in Section 3.

5 Conclusion

This paper presented a method for measuring the readability of Japanese texts using levelled corpora. First, we built a set of six-level corpora using text data extracted from textbooks of Japanese and National Diet meeting transcripts. We examined these corpora both manually and statistically. Then a multiple regression analysis of the results of these examinations was carried out. Among five models produced, we selected the best one and used it to construct our readability formula. The formula was tested using another set of levelled corpora built from 25 years of JLPT tests, and its reliability was confirmed. Our readability assessment formula is original in that it is built upon corpora of textbooks for learners of Japanese as a foreign language and thus it is considered more usable to assess the readability of texts used to teach or learn Japanese than other formulae developed on corpora of texts written for native readers of Japanese.

Moreover, we developed a web-based system using the formula to aid teachers of Japanese in preparing reading materials that match the level of their students. The system is also equipped with many reading-related functionalities that make it helpful not only to teachers, but also learners. Text highlighting according to the fine-grained six-level vocabulary list and pop-up dictionary with word definitions and example sentences are among the functionalities developed especially having learners' convenience in mind. Although a few limitations exist in this system, it is hoped that the system will enable a wide range of people involved in Japanese language instruction to benefit from the present research.

Literature

- Den, Y. [伝康晴]. (2009) Tayō na mokuteki ni tekishita keitaiso kaiseki shisutemu yō denshika jisho [多様な目的に適した形態素解析システム用電子化辞書] (“A multi-purpose electronic dictionary for morphological analyzers”). *Journal of the Japanese Society for Artificial Intelligence* 24(5), 640-646.
- Flesch, R. (1948) A new readability yardstick. *Journal of Applied Psychology* 32(3), 221-233.
- Kawamura, Y. [川村よし子]. (1999) Goi chekkā o mochiita dokkai tekisuto no bunseki [語彙チェッカーを用いた読解テキストの分析] (“Analyzing text for reading using a vocabulary checker”). *Kōza Nihongo Kyōiku* [講座日本語教育] (“Lectures on Japanese Language Teaching”) 34, 1-22.
- Lee, J. [李在鎬]. (2011) Daikibo tesuto no dokkai mondai sakusei katei e no kōpasu riyō no kanōsei [大規模テストの読解問題作成過程へのコーパス利用の可能性] (“Using corpora to create materials for reading section of large-scale tests”). *Nihongo Kyōiku* [日本語教育] (“Journal of Japanese Language Teaching”) 148, 84-98.
- Sakai, Y. [酒井由紀子]. (2011) Kenkō igaku jōhō o tsutaeru nihongo tekisuto no rīdabiritī no kaizen to sono hyōka: Ippan muke shippeī setsumei tekisuto no yomiyasusa to naiyō rikai no shiyasusa no kaizen jikken [健康医学情報を伝える日本語テキストのリーダビリティの改善とその評価：一般市民向け疾病説明テキストの読みやすさと内容理解のしやすさの改善実験] (“Improvement and evaluation of readability of Japanese health information texts: An experiment on the ease of reading and understanding written texts on disease”). *Library and Information Science* 65, 1-35.
- Sakamoto, I. [阪本一郎]. (1964) Bun no nagasa no hijū no sokuteihō: Readability kenkyū no kokoromi [文の長さの比重の測定法：Readability 研究の試み] (“Assessing the weight of sentence length: An attempt to approach the readability”). *Dokusho Kagaku* [読書科学] (“Science of Reading”) 8(1), 1-6.

- Sato, S. [佐藤理史]. (2011) Kinkō kōpasu o kihan to suru tekisuto nan'ido sokutei [均衡コーパスを規範とするテキスト難易度測定] (“Measuring text readability based on balanced corpus”). *IPSJ Journal* 52(4), 1777-1789.
- Shibasaki, H. [柴崎秀子] and Shin-ichiro Hara [原信一郎]. (2010) 12 gakunen o nan'i shakudo to suru nihongo rīdabiriti hanteishiki [12学年を難易尺度とする日本語リーダビリティ判定式] (“The readability formula to predict school grades 1-12 based on Japanese language school textbooks”). *Keiryō Kokugogaku* [計量国語学] (“Mathematical Linguistics”) 27(6), 215-232.
- Smith, E. A. and J. P. Kincaid. (1970) Derivation and validation of the automated readability index for use with technical materials. *Human Factors* 12(5), 457-464.
- Sunakawa, Yuriko, J. Lee, and M. Takahara. (2012) The construction of a database to support the compilation of Japanese learners' dictionaries. *Acta Linguistica Asiatica* 2(2), 97-115.
- Tateishi, Y. Y. Ono, and H. Yamada. (1988) A computer readability formula of Japanese texts for machine scoring. *Proceedings of the 12th Conference on Computational Linguistics*, 649-654.
- Zhang, Y. and K. Ozeki (1998) Automatic *bunsetsu* segmentation of Japanese sentences using a classification tree. *Language, Information and Computation (Proceedings of PACLIC12)*, 230-235.

要旨 (Abstract in Japanese)

「日本語のレベル別コーパスに基づいたリーダビリティ測定」

李在鎬 (早稲田大学)、長谷部陽一郎 (同志社大学)

本研究ではコーパスを用いて日本語テキストのリーダビリティを測定する方法の開発を行った。これにあたり2種のレベル別コーパスを構築した。リーダビリティ測定用の公式を構築するためのモデルとなるコーパスと、得られた公式の妥当性・信頼性を評価するためのコーパスである。作業は次のように行われた。まず、日本語の教科書と国会議事録から抽出したデータから、6レベルのモデルコーパスを作成した。次に、回帰分析を用いて得られたモデルの中から最も予測精度の高いものを選び、それを元にリーダビリティ公式を構築した。次にこの公式を、25年分の旧日本語能力試験 (Japanese Language Proficiency Test; JLPT) の読解問題データから作成した評価用コーパスに適用した。その結果、この公式によって高い精度で日本語テキストのレベル判別が可能ながことが明らかとなった。現在、この成果を元にして開発したオンラインのリーダビリティ判定システムを公開している。