A Computer Readability Formula of Japanese Texts for Machine Scoring

TATEISI Yuka, ONO Yoshihiko, YAMADA Hisao

Department of Information Science, Faculty of Science, University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113, Japan

Abstract

A readability formula is obtained that can be used by computer programs for style checking of Japanese texts and need not syntactic or semantic information. The formula is derived as a linear combination of the surface characteristics of the text that are related to its readability: (1) the average number of characters per sentence, (2) for each type of characters (Roman alphabets, kanzis, hiraganas, katakanas), relative frequencies of runs (maximal strings) that consists only of that type of characters, (3) the average number of characters per each type of runs, and (4) *tooten* (comma) to *kuten* (period) ratio.

To find the proper weighting, principal component analysis (PCA) was applied to these characteristics taken from 77 sample texts.

We have found a component which is related to the readability. Its scores match to the empirical knowledges of reading ease. We have also obtained experimental confirmation that the component is an adequate measure for stylistic ease of reading, by the cloze procedure and by the examination on the average time taken to fill out one blank of the cloze texts.

1. Introduction

This study aims to obtain a readability formula that can be used by computer programs for style checking of Japanese texts. A readability formula predicts the difficulty of a document that may result from its writing style, but not from its content, organization, or format. A readability index is calculated from the measures of surface characteristics of the document that are thought to indicate the stylistic difficulty without an attempt to parse sentences or to consult a large dictionary.

Many of the readability formulae for English, (for example, Flesch's Reading Ease Score /Flesch 1949/ and Automated Readability Index /Smith 1970/), use the average length (number of syllables or letters) of words and the average number of words in sentences in a document for calculating the readability index. Word length is a measure of the lexical difficulty, i.e., difficulty of the vocabulary used in the document. Sentence length is a measure of the syntactic difficulty or complexity of the sentence. While readability indices are derived from simple formulae, they predict reasonably well the difficulty of a document. This is because the sentence length and the word length are highly correlated with features such as the complexity of the sentence and the difficulty of the word, respectively.

Existing scoring methods for Japanese, such as the one proposed by /Morioka 1958/ or /Yasumoto 1983/, use the sentence length measured in letters instead of words and the percentage of *kanzis* (Chinese characters), the latter used for estimating the difficulty of the vocabulary. Both rate the average number of letters per sentence and the percentage of kanzis in the text independently and do not combine the two factors into a single index. A text with longer sentences is estimated as difficult, and a text with more kanzis is also estimated as difficult. Morioka, who surveyed on school textbooks, showed that the upper grade textbooks contain longer sentences on the average and more kanzi. Yasumoto states that documents with more kanzi are less readable even for adults, for the following reason. Kanzi are logograms, one roughly corresponding to a word. Documents using more kanzis, therefore, apt to include more different words and should demand more reading skill.

A problem of rating the sentence length and the percentage of kanzi independently is that these two may yield an inconsistent rating. Generally, a sentence becomes longer if its kanzis are rewritten in kanas. Thus sentence lengths depend on representations. There seems to have been no attempt on combining the factors of sentence length and the proportion of kanzi. On the other hand, no rationale is given for the separate measurements. It is possible to derive a single index that can assess readability of Japanese text.

/Sakamoto 1967/ proposed a method of scoring the relative difficulty of children's books to match the reading skill of the intended readers. His method consists of three independent ratings; (1) the proportion of fundamental words based on /Sakamoto 1958/, (2) the proportion of sentences that are made of more than 10 words, and (3) the proportion of kanzi. However, Sakamoto's method introduce the problem of measuring sentence length in words in place of the conflict between sentence length and representation.

Using word count or word length as an estimator of readability is not practical in the case of Japanese. Since Japanese does not use word segmentations in normal writing, dividing sentences into words needs parsing and consulting dictionary. Thus, a scoring method based on words, such as Sakamoto's, is costly. This is especially so when scoring is done by a computer, because extra devices such as parsers, a large dictionary, and, sometimes, semantic analyzers are required for word segmentation alone.

Another problem with the traditional scoring methods is that they have ignored katakana, which are used to represent foreign words. Recent documents, especially scientific and technical ones, use a lot of foreign words. /Watanabe 1983/ reports that, in a year's issues of the Journal of Information Processing Society of Japan, Vol. 17, about an eighth among the characters used is katakana. /Satake 1982/ surveyed the article of magazines published today and found that the ratio of katakana ranged from 4.44 to 13.75 percent. Thus percentage of katakana is not negligible in scoring today's documents. Katakana words mean imported foreign words, old and new, which are often unfamiliar to readers. Yet existing measures take into account only kanzi and are insufficient to score the today's technically oriented documents.

2. Factors of Readability

We have chosen the following four surface characteristics as factors of readability:

(1) relative frequency of characters for each type of characters,

(2) the length of a *run* (maximal string that consists of one type of characters),

(3) the length of a sentence, and

(4) the number of *tooten* s (commas) per sentence.

The former two are related to the difficulty of vocabulary in a document; the latter two are related to the complexity of sentences in a document.

Character Frequencies

The most common Japanese writing system is based on the mixture of *kanzis*, *kanas* (*hiraganas* and *katakanas*), the Roman alphabets, Arabic numerals, and some other alphabets and symbols. Almost all normal writing is a mixture of kanzis, hiraganas, and katakanas (and others). Frequencies of types of characters in a Japanese text are known to affect its readability at least in the following manner: Kanzi, as mentioned before, are considered to make texts difficult. Since katakana and alphabets are used for foreign words, high frequencies of these characters indicate that the text contain many unfamiliar words. Hiragana are used to represent the rest of the text and more of them are considered to make texts easier. There is no rigid orthography for Japanese. Nevertheless, the way an adult Japanese spells out a sentence in usual writing is roughly fixed. Kanzis are used for nouns and for the root parts of verbs, adjectives, adverbs, and the like. Hiraganas are used to write inflections and other grammatical parts of sentences, and katakanas are used mainly for the transcription of foreign words. So in passages written in the common way, the use of types of characters, i.e., kanzi, hiragana, katakana, etc., reflects the use of vocabulary and can be an indicator of the difficulty of the passage.

It is possible to write the words usually written in kanzi in hiragana. However, psychological experiments such as the ones conducted by /Kitao 1960/ or /Hirose 1983/ a reader finds it difficult to read the texts represented in the way unfamiliar to the reader. In Kitao's experiment, subjects took less time to read and recognize the word or the sentence written in a common way than written solely in hiragana. In Hirose's experiment, the words usually written in kanzi are harder to recognize than the words usually written in kana when both type of words are written in kana. Both results show that words or sentences in the representation more familiar to a reader are more readable than those in less familiar representation.

Runs

In the ordinary representation, a boundary of the types of characters corresponds to the boundary of words or smaller grammatical parts thereof. That is, a series of letters of the same type in the text, bounded by other character types corresponds to a word or a smaller grammatical part. We will call such a series a *run*, i.e., a run is a maximal string that consists of only one type of characters. It is not a grammatical unit. Usually, a run corresponds to one or more words. A verb or an adjective is often found across two runs. Such a word normally has its root part written in kanzi and its inflection part in hiragana.

As the boundary of runs roughly correspond to the boundary of words, the different graphic appearance of kanzi and kana letters helps a reader to parse a sentence. Hence, long runs, when they happen, hide the word boundaries and makes a sentence less readable.

Long kanzi runs give another problem to the readability. Kango can be formed into a compound word simply by concatenating two or more of them successively. The meaning of the new word is formed by the meanings of its elements. However, how each element is related to each other in the compound word is not clear from mere concatenation. A reader must pragmatically *see* the relation. Therefore, it is often the case that the meaning of a compound kango is ambiguous. For example, *siken-ki* can be read as *siken-suru-kikai* (testing machine) or as *siken-sareru-kikai* (machine to be tested); *rinzi-kyouiku-singi-kai* meaning *rinzi-ni-kyouiku-ni-tuite-singi-suru-kai* (an ad hoc council to deliberate on education) can be read as *rinzi-no-kyouiku-ni-tuite-singi-suru-kai* (a council to deliberate on an ad hoc education).

It is unlikely that there may be any good theory possible about the relationship between run frequencies and readability. Nevertheless, the run frequencies may be used in a similar manner as character frequencies. In a study preceding this /Tateisi 1987/ we found that the run frequencies are correlated with the frequencies of the character of corresponding types ($0.6 \leq r \leq 0.9$, depending on character types) and a unit of run is sufficient to obtain the information otherwise supplied by both characters and runs.

Sentence Length

The length of sentences is a known factor of readability as /Morioka 1958/ and other surveys show. In Japanese, as in other languages, long sentences tend to have complicated structures.

Sentence length can be measured in the number of characters it contains. Though Sakamoto's survey of children's textbook /Sakamoto 1963/ shows that the number of words per sentences is a more accurate indicator of the grade level than the number of characters, it also shows that the two are in good proportion, the correlation coefficient being 1.00.

Punctuation

Tootens, like commas, are put at the end of a phrase. The number of tootens per sentence corresponds to the number of phrases per sentence. /Hayasi 1959/ found that junior high school students and senior high students understood the text more precisely if modifying phrases are separated and made into independent sentences. Following this result, a sentence with smaller number of phrase is easier to understand. /Kozuru 1987/ found that the average number of tootens in a sentence increases with student's grade level. These findings indicate that the number of tootens in a sentence is greater in more difficult-to-read texts. Thus, the number of tootens is a factor of readability.

3. The Method of Analysis

We shall first extract several numerical characteristics of style from texts and then derive a readability formula as a linear combination of the values of those characteristics. A numerical index is only a rough scale of readability. It should be calculated with simple devices and methods. We use character as the unit of measuring length for the sake of simple calculation.

Several surface characteristics are extracted from the materials. Difference of the characteristics among materials consists of several factors. It may be factored into variation of the topic area of the texts, and the variation of style. Style may differ by the writer or by the intentions of the text. Introductory textbooks should be written easier than technical papers intended for experts and the authors will be careful not to make it difficult to read. Thus they will be written in a style easier to read than the style of technical papers. Translations tend to have a particular style, highly dependent on the syntax of the original language. The particular style of translations is often found awkward as Japanese and less readable. The distinctive feature of the texts with different intentions can be used as a criteria of assessing readability.

To find the distinctive feature of texts from the surface characteristics, the principal component analysis (PCA) extracts factors of variance of the characteristics. We will then examine the components, by comparing component scores for the materials with the empirical knowledges of readability. In this way we shall choose a component relevant to the stylistic readability. A principal component is a linear combination of the variables. The formula which computes the component can be used as a readability formula.

Variables

We have chosen the ten variables that represent the four factors of readability:

(1) for each type of characters((Roman) alphabets, kanzis, hiraganas, katakanas), relative frequency of runs (maximal strings) that consists only of that type of characters,

(2) the average number of letters per each type of runs,

(3) the average number of letters per sentence, and

(4) *tooten* to *kuten* ratio.

Sentence length is measured in the number of characters between two adjacent sentence-ending marks (*kuten*, exclamation marks, and question marks). Kuten, unlike period, is placed only at the end of a sentence, not as an indicator of abbreviations. Therefore, the end of a sentence is almost always detected by detecting kuten, although the end quotation embedded in a sentence is also counted as the end of a sentence.

Samples

We must compare the readability among the texts written in the common way, that is, the texts written by authors as they are. For example, the textbooks for elementary school children are inadequate. This is because those textbooks are written in an unusual way. They use hiragana where most adults use kanzi, transcribing the kanzi the readers are not expected to learn yet.

We will therefore take the documents written by adults for adults as materials of the analysis.

Seventy-seven (77) documents were selected as sample texts to extract the data from. Seventy of the samples are machine-readable documents that were stored in our laboratory. They are technical papers, textbooks for collage students, and translations of computer science materials, written by 13 authors. Seven of the samples are included as indicators for reading ease. Five of these indicators are text judged as easy. Three of them are taken from the books on technical writing; two are taken from essays for general readers. They are considered to be easier than the papers or textbooks for scientists. The remaining two are the text judged as difficult. One of them is a decision on the case of an infringement of copyright of a computer program; the other is a juridical paper about copyright and new media such as magnetic tapes. Juridical texts are empirically known as hard to read.

Tables, figures, references, and expressions which are displayed independently from the passage are deleted from the samples.

4. Result of the Principal Component Analysis (PCA)

The principal component analysis is done by S routines /Becker 1984/ on Vax 8600 at the Computer Center of the University of Tokyo. The components and the loadings of each variables are shown in table 4–1.

The first three components (eigenvalue > 1) are examined. Total variance explained by these components is 70%. Figure 4–1 shows the scatter plot of sample texts. The letter *i* designate introductory textbooks, *m* magazine articles other than technical papers, *p* technical papers, *t* and *T* designate translations from English papers, and *D* and *E* designate the difficult and easy indicators, respectively.

The following are observed for the first component.

(1-1) This component reflects the occurrences of alphabets; separates the texts with little alphabetic content and the text abundant with alphabetic content.

(1-2) The texts with many equations and abbreviations have high scores on this component.

The score on this component shows the area of topic.

The following are observed for the second component.

(2-1) This component separates the texts with long sentences and long kanzi runs from the other texts.

(2-2) The component score agrees with human judgement about easy/difficult texts. It is high on the texts judged easy and low on the texts judged difficult. The second component score shows the distinction more clearly than the first or the third.

(2-3) Introductory textbooks have generally higher scores than papers. Again, the second component score shows the distinction more clearly than the first or the third.

Since long sentences and long kanzi runs make texts less readable as stated before, (2-1) indicates that the second component can be an indicator of readability. (2-2) and (2-3) also indicates that the second component is related to readability.

The third component shows a difference of proportions of katakana and kanzi. From table 4–1 we can find that the variables on kanzi have positive loadings and the variables on hiragana and katakana have negative loadings on the component. Thus, the component shows the proportion of kanzi, in the way that it increases with texts with more kanzi.

5. Principal Component Scores and Style

We have observed the following phenomena on the second component.

Improvement and Principal Component Scores

Five of the sample texts are chapters (indicated *T* in the figure 4–1) of the final versions of the translation of an English paper by different translators. Their component scores were compared with those of the respective draft versions. (The drafts are not among the samples.) The first three component scores of the final manuscripts were uniformly higher than those of drafts, i.e., the scores became higher with the improvement of their style. The differences between the final versions and the respective draft versions are shown in table 5–1. The mean difference of the second component is found greater than that of the first at the 5 percent significance ($p = 0.044$) and greater than that of the third at the 10 percent significance but not at the 5 percent significance ($p = 0.098$). Thus, the difference of the second component is greater than the other two. This agrees with the observations on the distribution of texts, that is, easier-to-read texts have higher second component score than difficult ones, since a text becomes easier to read after improvement in general.

Frequencies of Passive Forms

Table 5–2 below shows the correlation between the component scores and the frequencies of passive. Passive forms are counted using the pattern matching method proposed by /Ushijima 1987/. The count is divided by the number of the kutens in a sample, yielding the ratio to passives per sentences, or sentence-endings.

Japanese passive forms are also used for potentials. For example, *mirareru* may mean either *be seen* (passive) or *can see* (potential) and *taberareru* may have one of three meanings: *be eaten, can eat,* and *can be eaten.* Thus, frequent use of passives tend to make a document vague and less readable.

The second component scores have a higher correlation than other component scores. Note that the correlation coefficient is negative. This agrees with the observation that the second component score is lower on difficult-to-read texts and that the frequency of passives is higher on such texts.

Figure 5–1 shows the plot of the second component scores and the frequencies of passives per 1000 sentences. The line in the figure is the regression line.

6. The Derived Formula

The results above support the adequacy of the second component as a scale of readability. To summarize, the second component score may be used as a readability index because of the following facts.

(1) The component score agrees with human judgement about easy/difficult texts. Easier-to-read texts yield higher valued scores.

(2) Introductory materials give higher scores than technical papers,

(3) The score increases as the result of improvement by editing of texts,

(4) The frequencies of passive forms have a negative correlation (-0.53) with the component score.

The first component and the third component do not possess all of these properties. Thus the second is a better measure of readability than the first or the third.

The second component score is transformed so that the mean on those 77 samples equals 50, the standard deviation equals 10, and let the value be higher on easy texts. This yields formula,

$$\begin{aligned} RS &= 0.06 \times pa + 0.25 \times ph - 0.19 \times pc - 0.61 \times pk \\ &\quad -1.34 \times ls - 1.35 \times la + 7.52 \times lh - 22.1 \times lc - 5.3 \times lk \\ &\quad -3.87 \times cp - 109.1 \end{aligned}$$

where pa, ph, pc, pk are the percentages of alphabet runs, hiragana runs, kanzi runs, and katakana runs, respectively; ls is the average numbers of letters per sentence; la, lh, lc, lk are the average numbers of letters per alphabet run, hiragana run, kanzi run, and katakana run, respectively; and cp is the tooten to kuten ratio.

7. Validation of the Derived Formula

We have also obtained experimental conformation on the idea that the RS is an adequate measure for stylistic ease of reading, by the cloze procedure /Taylor 1953/, /Shiba 1957/.

Cloze procedure judges the relative reading difficulty of texts to a particular population. This difficulty mostly related to the content of the text. Suppose readers have no background knowledge of the content. They are not likely to be able to fill a blanked-out word where a technical term or some other word that requires the knowledge of the area the content belongs to be filled in. In such cases, the cloze score or the cloze percentage becomes low for the text even if an experts finds it very easy to read.

Stylistic difficulty may be also measured by this procedure, according to the experiment of /Kitao 1960/. In the experiment subsequent to the one mentioned in the previous section, he required the subjects to perform the cloze procedure on two materials; the same text represented in two different ways. One is in the usual representation, mixing kanzi and kana; the other is written entirely in hiragana. The cloze score of the usual form was higher than the one entirely in kana. This result was consistent with the result that the subjects required longer time in reading the text entirely in kana, as mentioned in section 2.

As the cloze procedure scores both the difficulty of style and the difficulty of the content, another measure is needed to confirm that our formula is a measure of the stylistic readability. For this purpose, we recorded the total time each subject took to complete a cloze text. The recorded time was divided by the number of blanks, thus converted into the average time taken to fill out one blank.

The process a subject takes to fill out a blank is composed of four phases, i.e.,

(1) the phase of reading the incomplete text and understand the content of the passage,

(2) the phase of surmising what is missing (as a notion),

(3) the phase of choosing the proper word to supplement, and

(4) the phase of writing down that word.

The time for writing down a word is fairly constant, unless the word contains extremely complicated kanzi. Therefore, the variation of time from text to text is the variation of time for the phase (1), (2), and (3), i.e., understanding the passage, surmising the missing notion, and choosing the proper word. The text which is stylistically difficult takes more time in the phase (1). Thus the difficult-to-read texts must require more time filling out blanks than easy-to-read ones.

Materials

The materials of the experiment, denoted by $p\,1$ through $p\,6$, were taken from the six sample papers among the 77 used for the PCA. Each was about 500 characters in size. Three of them ($p\,1$, $p\,2$, $p\,3$) had high RSs ($RS > 50$) and the three ($p\,4$, $p\,5$, $a\,6$) had low RSs ($RS < 50$).

Every eighth word of each text was blanked out, i.e., the proportion of blanked out words to the whole words was 12.5%. Ten underscore ('_') characters were put where a word was blanked out. Among several different definitions of Japanese words used, the one which gives the smallest unit was taken. The materials were printed out on a sheet of A4-sized paper, one material per paper.

Twenty-eight subjects (25 undergraduate students and 3 graduate students) participated in the experiment. Each subject was assigned three materials from $p\,1 - p\,6$ selected randomly, so that the half of the subjects were assigned each material. The subjects were required to fill blanks (underscored parts) with words they thought most appropriate to the context, taking as much time they need. The subjects were told that each word-unit was smallest possible, and therefore the deleted part might not match what they think is a word. The subjects were also told that the materials that they had were independent from each other.

At the same time, the subjects were required to record the time when he/she start to fill out each paper, i.e., one material, and when he/she completed, for each paper, to the unit of seconds.

Results

Completed sheets, expect for one by a subject who gave up the procedure in the middle are analyzed. Whether the word filled in matched the original or not was judged according to /Shiba 1957/. Some sheets are without the record of the time. Such sheets are included for calculation of cloze percentages but excluded from the analysis of time. The cloze percentages and the medians of the time taken to fill a blank are shown in table 7–1. The cloze percentages were higher on texts with higher RS, although the correlation was not statistically significant (the correlation coefficient between cloze percentages and RSs is 0.295).

For the analysis of time taken, the texts were divided into two categories; the ones with $RS > 50$ and the ones with $RS < 50$. The average time for filling a blank was compared between the two categories of the texts using the median test. The result is shown in table 7–2. The difference of the time for filling a blank is shorter on texts with high RSs.

In addition, we compared a document and its rewritten version by the same procedure. The material $r\,1$ was taken from the final report of *rinzi-kyouiku-singikai* (National Council of Educational Reform) of Japanese Government. The document $r\,1$ had an extremely low score ($RS = 27$). The material $r\,2$ was rewritten from $r1$ by dividing long sentences into shorter sentences and substituting Japanese words for words of Chinese origin. The intention of the rewriting is to increase *RS*. The *RS* of the rewritten text is 47, nearly the average of the sample texts of PCA (= 50). The cloze percentage and the average time for filling a blank is compared compared as above. The cloze percentage of the rewritten version $r2$ was 59.6 %, higher than that of the original $r1$ (56.6 %). The average time for billing a blank is shorter for the rewritten version than for the original (the median was 9.6 sec. for $r2$ and 10.9 sec. for $r1$). The median test of time comparing the two materials showed that the difference was not statistically significant ($\chi^2 = 0$).

These results show that

(1) the subjects take shorter time with the texts of higher *RS* in understanding and guessing the missing words of the text, than with the text of lower *RS*, though the result is not statistically significant, and that

(2) the subjects guess the missing words in the high *RS* texts more correctly than those in the low *RS* texts.

The difference in time a subject spent to fill one blank in the two types of texts is significant, by the median test.

These results did not show that *RS* is related to the difficulty of the content or the vocabulary of the texts. However, *RS* is related to the stylistic difficulty, that is, *RS* show the relative difficulty of transformation from the text itself to the content. Therefore, *RS* is judged useful to measure the readability of texts in general.

We judged that the cloze score is more related to the difficulty of the content than to the difficulty of the style. Therefore we introduced another measure for stylistic difficulty. A comment from the experimental subject who gave up the procedure confirms our judgement to be reasonable. He gave up the task because he has not enough knowledge of the area of these texts, especially, of technical papers.

8. Concluding Remarks

We have derived a readability formula from the multivariate analysis on variance of surface characteristics of Japanese technical documents intended for adult readers.

The mean, the minimum, and the maximum value of RS over the several types of texts are shown in table 8–1.

As with all indices, RS can be increased by revision which does not necessarily enhance readability. For example, if a text is written entirely in hiragana and the sentences are cut into short shorter ones, *lh* and *ph* increases and *ls* decreases. This revision yields greater value of RS but does not produce the text easier to read. The formula should be applied to the texts written in the common way. To construct an index that is sensitive to the unreadability caused by unusually many hiraganas, we may need a quadratic formula on hiragana run length or hiragana run frequencies.

References

/Becker 1984/ Becker, R. A. and Chambers, J. M., "S: An Interactive Environment for Data Analysis and Graphics", Wadsworth,Belmont, California, 1984

/Flesch 1949/ Flesch, R., "The Art of Readable Writing", Harper,1949

/Hayasi 1959/ Hayasi, S., "*Yomi no Nooryoku to Yomiyasusa no Yooin to Yomareta Kekka to*" Mathematial Linguistics, Vol. 11, pp.20-33, 1959 (In Japanese)

/Hirose 1983/ Hirose, T., "The Effect of Script Frequency on Semantic Processing of Kanji and Kana Words", Jap. J. of Psychol., Vol. 55, No. 3, pp. 173-176, 1984 (In Japanese)

/Kitao 1960/ Kitao, N., "Comparative Study on Readability of'Hiragana-bun' and 'Kanji-majiri-bun'", Jap. J. of Educ. Psychol., Vol.7, No. 4, pp.1-5, 1960 (In Japanese)

/Kozuru 1987/ Kozuru, Y., "Basic Study for Readability Estimation of Japanese Documents", Proc. of the 34th Convention of IPSJ, pp. 1295-1296, 1987 (In Japanese)

/Morioka 1958/ Morioka, K., "Readability". In: Endo M. (Ed), "*Kotoba no Kagaku*", *Nakayama Shoten*, Tokyo, 1958 (In Japanese)

/Sakamoto 1958/ Sakamoto, I., "*Kyooiku Kihon Goi*", *Gakugei--Tosyo* ,Tokyo, 1958 (In Japanese)

/Sakamoto 1963/ Sakamoto, I., "Assessing the Weight of SentenceLength --- An attempt to Approach the Readability" Science of Reading, 7, pp. 1-6, 1963 (In Japanese)

/Sakamoto 1967/ Sakamoto, I., "A Yardstick for Readability", Science of Reading, 14, pp. 1-6, 1967 (In Japanese)

/Satake 1982/ Satake, H., "On the Frequency Ratio of Kinds of Letters in All Sorts of Sentence", Report of The National Language Research Institute No. 71, p.p.327- 346, 1982 (In Japanes)

/Shiba 1957/ Shiba, S., "A study of Readability Measurement --- Application of Cloze Procedure to Japanese Language" Jap. J. of Psychol., Vol.28 No.2, pp.67- 73, 1957 (In Japanese)

/Smith 1970/ Smith, E. A. and Kinkaid, P., "Derivation and Validation of the Automated Readability Index for Use with Technical Materials", Human Factors, Vol. 12, pp. 457-464, 1970

/Tateisi 1987/ Tateisi, Y., Ono, Y., and Yamada, H., "Statistical Analysis of Japanese Texts as a Basic Stuby for Readability", Proc. of the 3rd Symposium on Human Interface, pp. 15-22, Osaka, 1987 (In Japanese)

/Taylor 1953/ Taylor, W. L., "Cloze Procedure: A New Tool for Measuring readability", Journalism Quarterly, Fall 1953

/Ushijima 1987/ Ushijima, K., Ishida, M., Yoon J., and Takagi T., "A Simple Method to Extract Passive Voices in the Writing Tools for Japanese Documents", Trans. of IPSJ, Vol.28, No. 8, 1987 (In Japanese)

/Watanabe 1983/ Watanabe S. and Ogisi H., "*Zyoho Syori no Yozi to Yogo*", Preprint of Working Group WGJI 10-2, Information Processing Society of Japan, 1983 (In Japanese)

/Yasumoto 1983/ Yasumoto B., "*Settoku no Bunsyo Gizyutu*", *Kodan--sya* , Tokyo, 1983 (In Japanese)

Table 4–1. Component Loadings

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
Alpha. r. f.	0.87	0.03	0.03	-0.04	0.17	-0.39	0.10	-0.04	-0.22	-0.05
Hira. r. f.	-0.93	0.19	-0.13	0.03	-0.03	0.04	0.11	0.10	-0.22	0.11
Kanzi r. f.	-0.92	-0.14	0.24	0.0	-0.18	0.10	0.09	0.08	-0.08	-0.15
Kata. r. f.	0.01	-0.25	-0.85	-0.26	-0.11	-0.02	-0.33	0.12	-0.04	-0.03
Sent. length	-0.72	-0.34	-0.10	-0.05	-0.04	-0.55	0.16	0.07	0.13	0.02
Alpha. r. l.	0.34	-0.37	0.04	0.75	-0.39	-0.07	-0.12	0.06	-0.03	0.01
Hira. r. l.	-0.63	0.54	-0.22	0.25	0.02	-0.14	-0.19	-0.38	-0.01	-0.02
Kanzi r. l.	0.0	-0.78	0.25	-0.39	-0.30	0.02	-0.04	-0.28	-0.06	0.04
Kata. r. l.	-0.04	-0.63	-0.53	0.28	0.29	0.20	0.32	-0.14	0.0	-0.02
Tooten per Kuten	-0.43	-0.54	0.36	0.13	0.50	-0.03	-0.35	0.05	-0.03	0.01
Eigenvalue	3.66	1.95	1.34	0.95	0.65	0.53	0.45	0.29	0.13	0.04
Proportion (%)	36.60	19.50	13.40	9.50	6.50	5.30	4.50	2.90	1.30	0.40
Cumulative (%)	36.60	56.10	69.50	79.00	85.60	90.90	95.40	98.30	99.60	100.00

r. f. = run frequency, r. l. = run length

Fig. 4–1. Principal Component Scores

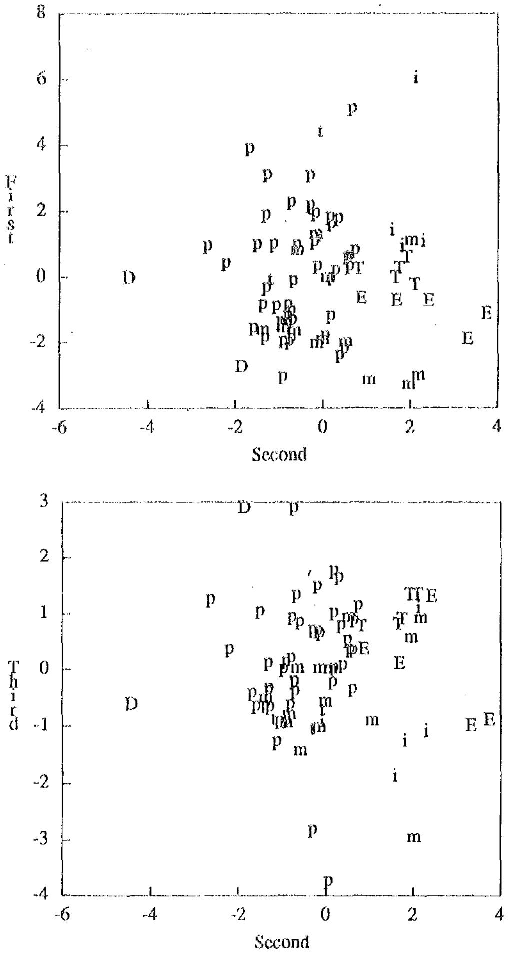

Table. 5–1. Score Change by Improvement			
	No. 1	No. 2	No. 3
text 1	0.47	1.18	0.34
text 2	0.11	0.60	0.87
text 3	0.41	0.38	0.23
text 4	0.0	0.42	0.20
text 5	.0.50	1.33	0.43
mean	0.30	0.78	0.41
sdev	0.10	0.20	0.12

Fig. 5–1. Frequencies of Passives

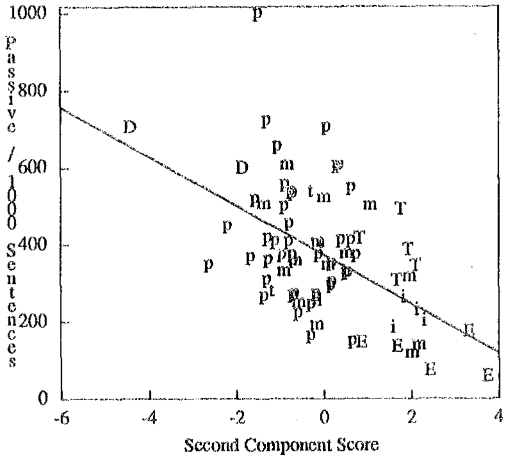

Table 7–1. Materials of the Experiment						
	p1	p2	p3	p4	p5	p6
RS	61.44	54.54	51.86	43.72	37.87	35.64
cloze %	66.58	63.81	56.28	56.25	64.69	60.46
time/blank (sec.)	6.89	6.19	8.18	8.75	9.06	8.05

Table 7–2. median test on time		
	$RS > 50$	$RS < 50$
long	12	24
short	24	12
$\chi^2 = 6.72, p < 0.05$		

Table 8–1. RS Values				
Text Type		Max	Mean	Min
PCA Samples	*Easy* Indicators	76.8	67.2	56.2
	Difficult Indicators	36.7	27.5	18.3
	Technical Documents	66.4	49.4	31.2
Textbooks	Junior High School	59.9	55.2	48.5
	Senior High School	58.0	49.2	39.5

In table 8–1, *indicators* are the five and the two texts in the PCA samples included as indicators judged as easy and as difficult. *Technical Documents* are the other 70 samples. *Textbooks* are the passages taken from the school textbooks on natural science and from the ones on social science, five for each. They are included in table 8–1 for comparison.

Table 5–2. Correlations to the Frequency of Passive Forms										
	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
r	-0.25	-0.53	0.02	-0.15	0.09	-0.06	-0.20	-0.01	-0.09	0.03