A Useful-based Evaluation of Reading Support Systems: Comprehension, Reading Speed and Effective Speed

Katsunori KOTANI
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan, 619-0289
kat@khn.nict.go.jp

Takehiko YOSHIMI
Ryukoku University
1-5, Yokotani, Setaoe-cho, Otsu-shi, Shiga, Japan, 520-2195

Takeshi KUTSUMI
Sharp Corporation
492 Minosho-cho, Yamatokoriyama-shi, Nara, Japan, 639-1186

Ichiko SATA
Sharp Corporation
492 Minosho-cho, Yamatokoriyama-shi, Nara, Japan, 639-1186

Hitoshi ISAHARA
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan, 619-0289

Abstract

This paper reports the result of our experiment, the aim of which is to examine the efficiency of reading support systems such as a sentence-machine translation system, a word-machine translation system, and so on. Our evaluation method used in the experiment is able to handle the different reading support systems by assessing the usability of the systems, i.e., comprehension, reading speed, and effective speed. The result shows that the reading-speed procedure is able to evaluate the support systems as well as the comprehension-based procedure proposed by Ohguro (1993) and Fuji et al. (2001).

1 Introduction

When one tries to read a passage written in a language, which is not familiar with him/her, one can choose a sentence-machine translation system (henceforth, S-MT system) and/or a word-machine translation system (henceforth, W-MT system). In choosing the appropriate system, evaluation of these systems may be in order.

Although there have been proposed various types of evaluation methods, e.g., BLEU (Papineni et al. 2002), these cannot evaluate an S-MT system and a W-MT system at once, nor compare these different systems. This is because they evaluate a system based on overall accuracy of system outputs. For instance, BLEU compares sentence-translation outputs with the reference translation. Thus, it is not easy to handle an S-MT system and a W-MT system at once.

In order to evaluate different systems, we determined to evaluate a system based not on the accuracy of outputs but on the usability. As some researchers (Ohguro 1993; Fuji 2001) suggest, an S-MT system and a W-MT system can be evaluated by examining comprehension test scores.

Along with these previous studies, we carried out an experiment, in which we examine the usability of some reading support systems from viewpoint of not only comprehension but also speed. The experiment results revealed that the efficiency of reading support systems might vary depending on the readers' ability.

This paper is organized as follows. Section 2 reviews the previous studies, which evaluated reading support systems based on the comprehension performance, i.e., Ohguro (1993) and Fuji et al. (2001). Section 3 describes our evaluation method, which evaluates several types of reading support systems based on the reading efficiency. Section 4 describes the experimental design. Section 5 reports the results of our experiment, in which we confirmed that the efficiency of systems might be different among the users' reading ability. Finally, section 6 presents our conclusions, and points out our future work.

2 Comprehension-based Methods

In this section, we will review evaluation methods, which applied comprehension test scores as a criterion in assessing the performance of reading support systems such as a W-MT system (Ohguro 1993) and an S-MT system (Fuji et al. 2001). These studies carried out experiments in which participants read both passages annotated

with translations and plain passages, and answered the relevant comprehension questions. On the basis of the test scores, these studies tried to evaluate the usability of these reading support systems.

In the following subsection, we will review Ohguro's (1993) study, which held on an evaluation experiment for a W-MT system. In the section 2.2, we will introduce Fuji et al.'s (2001) study that attested the usability of S-MT systems.

2.1 Ohguro (1993)

Ohguro (1993) examined the efficiency of an English-Japanese W-MT system, and suggested that a W-MT system would be more useful for those who had low reading ability.

Fifty-four non-native speakers of English took part in the experiment. The designed task was to read passages taken out of TOEIC reading section and to answer the relevant comprehension questions. Twenty-eight different passages and eighty comprehension questions were prepared, and assigned to each participant.

The experiment procedure was divided into two phases. First, all the participants read fourteen English-only texts, and answered the forty questions. Based on the test score, the participants were classified into two groups so as to balance the reading ability between them. Then, English-only texts and questions were given to one group, i.e., a control group, and texts annotated with word-translation were provided with the other group, i.e., experimental group.

Ohguro (1993) hypothesized that the test score of the control group would not be different from the score of the first test, since both of the tests were written only in English. By contrast, it was supposed that the score of the experimental group would increase because of the annotated word translation. It was further hypothesized that the increase of test score would depend on the reading ability, i.e., TOEIC scores (Hypothesis I). That is, the increase of the test score would be greater for the lower TOEIC score group than the higher TOEIC score group. This Hypothesis was not borne out. Then, the Hypothesis I was modified by changing the reading ability level from TOEIC score to the result of the first test. The revised hypothesis was verified. Thus, those who got low scores in the first test showed the greater increase in the second test scores than the higher test score group. On the basis of this experimental result, Ohguro (1993) concluded that the supporting effect of a W-MT system should be greater for less skilled readers than highly skilled readers.

2.2 Fuji et al. (2001)

Fuji et al. (2001) investigated the efficiency of English-Japanese S-MT systems, and suggested the similar effect to Ohguro's (1993).

Approximately two hundred non-native speakers of English participated in the experiment. The participants were divided into twelve groups based on the TOEIC score. The score range was below 395 to above 900. Each participant read fourteen texts, and answered forty comprehension questions. There were three types of texts. One was an English-only text, i.e., a control text, another contained only translation, and the final type of texts involved English texts annotated with sentence translation. Two different S-MT systems was used, and hence there were two subsets for translation-only texts and annotated English texts.

The experiment result did not show significant difference between the two S-MT systems, and thus, it was concluded that the performance of these systems should be similar to each other. The experiment results showed that the efficiency of S-MT systems would vary depending on the reading ability of participants. Thus, the higher TOEIC score group show less increase of score in translation-annotated tests than the lower score group, compared with the test score of English-only texts. In addition, while the score in translation-only texts increased in the case of the lower score group, the score for the higher score group decreased.

By examining the reading/answering completion time, Fuji et al. (2001) suggested that the MT-system should shorten the time for the lower score group to finish the test, unlike the higher score group.

2.3 Summary

Through the surveys of the previous studies, we found that reading support systems would show the greater efficiency for less skilled readers than highly skilled ones. Given this, we suppose that reading speed/reading efficiency evaluation should exhibit the similar supporting effect to these studies. That is, the less skilled participant would present the greater effect on the reading efficiency than the highly skilled group.

3 Evaluation with Reading Speed

In this section we briefly discuss why we adopt reading speed as another evaluation criterion. Then, we present the reading speed effect on readability of a text, which we found through our pilot study. Finally, we will describe our reading efficient-based evaluation procedures.

3.1 Reading Efficiency

In order to evaluate different reading support systems, we determined to examine to what extent these systems affect reading. We suppose that the appropriate supporting effect would increase reading efficiency from the baseline, i.e., text reading without any supports, as the previous studies (Ohguro 1993; Fuji et al. 2001) suggested it. Unlike these previous studies, we adopt reading speed as another criterion, because the reading efficiency should contain not only the comprehension performance but also the speed performance (cf. Alderson 2000).

We consider that the reading speed-based evaluation method takes several advantages over comprehension-only evaluation method. First of all, the application of reading speed as an evaluation criterion should more greatly explicate the effect of reading support systems than the comprehension-based method. This is because the reading speed-based method can be easily combined with the comprehension method. By examining the reading support systems from these viewpoints, the evaluation would depict the system performance in more detail.

Secondly, reading speed can be measured not only at the text level but also at the sentence level. Thus, the reading-speed method can evaluate the efficiency of a system even in this local domain. The comprehension-based evaluation should encounter some difficulty for such a local domain evaluation. For instance, each sentence should be accompanied with the relevant comprehension question. These questions apparently decrease the reading efficiency, and hence they should be avoided. The reading-speed method does not encounter this problem, and make it possible to evaluate the effect of systems at the sentence level. The examination of a system in the local domain effect makes it possible to investigate whether a system would be affected by some linguistic properties such as complexity of a syntactic structure, familiarity of words, and so on.

Finally, the reading-speed method can be applied to any types of texts as far as the reading support sytems can handle. The comprehension method, however, should be limited to texts involving relevant comprehension questions. Otherwise, one has to make appropriate questions, which would arise another problem, i.e., a cost problem. Thus, the speed-based method is able to attest the efficiency of reading support systems in any types of texts such as newspaper/magazine articles, web pages, emails, and so on, and so forth.

3.2 Reading Speed and Readability

Before using reading speed as an evaluation criterion, we carried out a pilot exmperiment, which exmined whether reading speed would reflect readability of a text to a certain degree. If the speed has nothing to do with the readability, the reading speed-based evaluation would not reveal the effect of reading support systems.

In the experiment, twenty participants read two English texts sentence-by-sentence, and the reading time for each sentence was measured. The reading time per sentence was divided by the number of words within a sentence, and the reading time data was changed into the reading speed (WPM) data.

In order to confirm whether the data could detect reading difficulties, the speed data was compared with readability scores calculated with Flesch's readability formula (Flesch 1948), which took into account both average word and sentence length. Examining the association of reading speed and readability scores found that there was statistically significant positive correlation between them ($r=0,7$, $p<0.01$). On the basis of the correlation, we concluded that the reading speed should reflect the readability of a text.

3.3 Reading Support System Effect on Readability

We assume that reading speed would reflect readability of a text, and we further suppose that the reading support systems should affect text readability. That is, a reading support system would decrease readability of a text, if the system appropriately functions. By contrast, when the performance of a system would be rather poor, the readability of a text would remain unchanged, or might increase.

Under our evaluation method, we suppose that the positive effect of a support system would increase the speed from the baseline, i.e., speed of reading texts without any support, while the negative effect should decrease it. Given this, examining reading speed of an English text and that of a supported text would reveal the effect of a system. Thus, the reading speed of less skilled readers would more readily increase than that of highly skilled readers as the previous studies (cf. Ohguro 1993; Fuji et al. 2001).

We further suppose that the effectiveness of reading speed (cf. Muter & Maurutto 1991) could be examined with reading speed weighted by comprehension rate, if comprehension test scores are available. The weighted reading speed, i.e., effective reading speed (eWPM), is derived from multiplication of reading speed by correct answer rate. The effective reading speed would exhibit the

similar effect to the other criteria such as comprehension test score and reading speed.

4 Experimental Design

A sentence machine translation (S-MT) system, a word machine translation (W-MT) system, and a chunker system were used. By using these reading support systems, four types of texts were prepared: English texts with slashes at the phrase boundary, machinery translated texts, English texts with sentence translations, and English texts annotated with word translations. In addition to these experimental texts, two types of texts were prepared as control texts: manually translated texts and English texts. These passages were taken out of TOEIC reading section questions in TOEIC preparation textbooks.

4.1 Participants

A hundred-three participants took part in this experiment. They ranged in TOEIC test score from 400 to 990, with a mean score of 690.43 (S.D. 164.94). All participants were English learners, and their native language was Japanese.

4.2 Task Design

Reading speed and comprehension were compared by having participants read sixty-eight passages. In order to calculate the reading speed, a reading-process monitoring tool, which enables to record reading time for each sentence, was used.

The reading-process monitoring tool showed a passage sentence-by-sentence on a computer screen, and recorded the time when a participant accessed each sentence (For more details of this tool, see Yoshimi et al. 2005). Thus, the participants read the texts in a self-paced reading fashion.

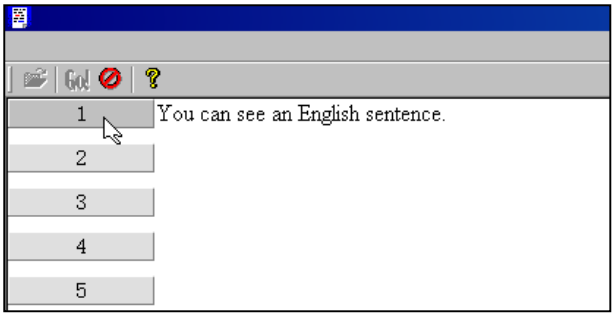

Figure 1: Screenshot of the Reading-Process Monitoring Tool

Each of the passages comprised of a text from one of the six text types (two controlled texts and four experimental texts), and contained some reading comprehension questions.

4.3 Procedure

Participants were instructed to read texts and answer the relevant questions. They were asked to read as quickly and as accurately as possible. The reading-process monitoring tool registered both reading time per sentence and the answers for the questions.

To accurately determine reading support system legibility and its associated effect on reading ability, reading speed or reading comprehension test score was used. The reading speed was derived from the number of words in a passage divided by the reading time taken to read the passage. The comprehension test score was calculated by dividing the number of questions with the number of correctly answered ones.

The reading ability of participants was determined with TOEIC score, and the participants were classified into six classes: (i) 400-495; (ii) 500-595; (iii) 600-695; (iv) 700-795; (v) 800-895; and (vi) 900-995. The mean reading speed for the passages was 82.45 WPM (S.D. 18.15), and the mean of the comprehension test scores was 0.80 (S.D. 0.10).

5 Results and Discussion

A within-classes ANOVA design was used to analyze objective differences between the text types. Post hoc comparisons were done by using the Dunnett test (95% C.I.).

Given this, we predict that our experimental result would show the similar supporting effect on comprehension and reading speed. We further suppose that the lower score group would present the greater increase of the reading efficiency than the higher score group.

5.1 General Result

On the basis of the TOEIC score of particpants, the participants were classified into six classes. The following table showed the mean test score, the mean reading speed, and the mean effective speed for each class. The higest mean score was found at 900-class, and the lowest at 400-class. By contrast, the mean speed was not straightforward to the classification, unlike the mean eSpeed.

Class	Comp.		Speed		eSpeed	
	M	S.D.	S.D.	S.D.	M	S.D.
400-	0.7	0.3	82	57	60	57
500-	0.7	0.3	79	56	61	57
600-	0.8	0.3	74	51	61	47
700-	0.9	0.2	84	54	73	56
800-	0.9	0.2	98	54	81	54
900-	0.9	0.2	92	51	82	51

Table 1: The Mean Comprehension Score, Reading Speed (WPM) & eSpeed (eWPM) Result

5.2 Reading Comprehension

Comprehension test scores were examined by comparing the scores between the controlled texts, i.e., manually translated texts (henceforth, JPN) and English texts (henceforth, ENG), and the experimental texts, i.e., English texts with slashes at the phrase boundary (henceforth, CHU), machinery translated texts (henceforth, MT-J), English texts with machinery translated sentences (henceforth, MT-EJ), and English texts annotated machinery translated words (henceforth, MT-W).

Examination of the mean test scores (0-1) for each text type found than the mean score of MT-EJ is significantly higher than that of ENG. The highest mean score was marked by JPN (Control Text). The second highest score was found in MT-EJ. The lowest score was seen in MT-W.

Features	Mean Score	S.D.
ENG(Control)	0.78	0.32
JPN(Control)	0.90	0.18
CHU	0.78	0.30
MT-EJ	0.83	0.25
MT-J	0.79	0.25
MT-W	0.74	0.32

Table 2: The Mean Comprehension Test Result

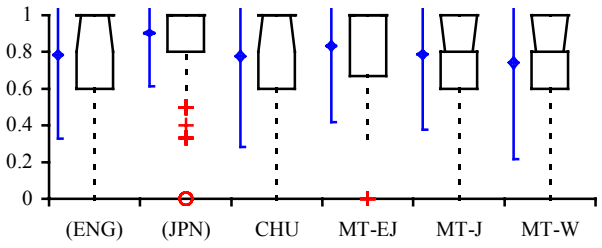

Figure 2: The Mean Comprehension Test Result

The following tables show the mean test scores for each text types. Table 3 presents the mean score of control texts, and Table 4 presents the mean score of experimental texts.

The mean score of ENG is fairly consistent with the ranking of TOEIC score. The mean score of JPN is also consistent with the ranking, although the difference is rather small.

Class	Text Type	
	ENG	JPN
400-	0.63	0.88
500-	0.67	0.88
600-	0.81	0.90
700-	0.86	0.94
800-	0.84	0.93
900-	0.95	0.91

Table 3: The Mean Test Score of Control Texts by TOEIC Score Group

The following table shows the mean test score of each text type for each TOEIC score group. If the score increases from the score of ENG in more than 0.05 point, an up-arrow appears. When the score decreases in more than 0.05 point, a downward arrow comes up. The crossbar indicates that the score change does not exceed 0.05 point.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	0.62 –	0.74 ↑	0.70 ↑	0.62 –
500-	0.68 –	0.74 ↑	0.77 ↑	0.67 –
600-	0.80 –	0.79 –	0.88 ↑	0.72 ↓
700-	0.82 –	0.80	0.86 –	0.80 ↓
800-	0.86 –	0.83 –	0.87 –	0.84 –
900-	0.90 –	0.84 ↓	0.92 –	0.85 ↓

Table 4: The Mean Test Score of Experimental Texts by TOEIC Score Group

By examining the mean test scores for each text type within the respective reading ability class, we found that none of the four experimental texts were read significantly more accurately than ENG in 400-class, 600-class, 700-class, and 800-class; MT-EJ texts were read significantly more accurately than ENG in 500-class; MT-J and MT-W texts were read significantly less accurately than ENG in 900-class.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	–	–	–	–
500-	–	–	↑	–
600-	–	–	–	–
700-	–	–	–	–
800-	–	–	–	–
900-	–	↓	–	↓

Table 5: The ANOVA Result of the Comprehension Test (ENG)

Further analyses revealed that all the four types of texts were read significantly less accurately than JPN texts in 400-class and 500 class; the texts other than MT-EJ were read significantly less accurately than JPN texts in 600-class and 700-class; none of the texts were read significantly less accurately than JPN texts in 800-class and 900-class.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	↓	↓	↓	↓
500-	↓	↓	↓	↓
600-	↓	↓	–	↓
700-	↓	↓	–	↓
800-	–	–	–	–
900-	–	–	–	–

Table 6: The ANOVA Result of the Comprehension Test Result (JPN)

5.3 Reading Speed

We examined reading speed by comparing the speed between the controlled texts, i.e., ENG and JPN, and the experimental texts. The mean speed of controlled texts was marked 62.39 WPM (S.D. 31.84) in ENG texts and 148.73 WPM (S.D. 72.02) in JPN texts. The ENG texts were read the most slowly, and the JPN texts were read the fastest.

Features	Mean Score	S.D.
CHU	63.97	33.44
MT-J	91.81	47.79
MT-EJ	64.63	31.84
MT-W	62.53	32.49
ENG(Control)	62.391	29.85
JPN(Control)	148.73	72.02

Table 7: The Mean Reading Speed

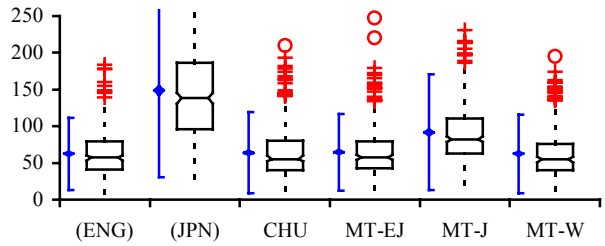

Figure 3: The Mean Reading Speed

The following tables show the mean reading speed for each text types. Table 8 presents the mean speed of control texts, and Table 9 presents the mean speed of experimental texts.

The mean speed of ENG is apparently inconsistent with the ranking of TOEIC score. If the score groups are composed of three classes, the speed turns out to be consistent with the ranking. That is, the speed for the novice group, i.e., 400-500 class, was 52.91 WPM, the speed of the intermediate class, i.e., 600-700 class, was 60.44 WPM, and the speed of the expert group, i.e., 800-900 class, was 75.45 WPM. The mean score of JPN seems invariant through the classes.

Class	Text Type	
	ENG	JPN
400-	55.25	147.03
500-	50.57	159.29
600-	61.00	135.17
700-	59.87	158.24
800-	68.87	147.90
900-	82.02	147.539

Table 8: The Mean Speed of Controlled Texts

The following table shows the mean speed of each text type for each TOEIC score group. If the increase of speed exceed 5 point compared with the speed of ENG, an up-arrow appears. When the speed decreases in more than 5 point, a downward arrow comes up. The crossbar indicates that the speed change does not exceed 5 point.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	52.89 –	99.68 ↑	78.37 ↑	58.63 –
500-	55.63 ↑	93.86 ↑	59.92 ↑	53.74 –
600-	55.11 ↓	90.61 ↑	51.36 ↓	50.11 ↓
700-	62.68 –	86.63 ↑	63.24 –	63.51 –
800-	82.94 ↑	97.25 ↑	74.70 ↑	83.92 ↑
900-	82.89 –	84.94 –	73.15 ↓	78.56 –

Table 9: The Mean Speed of Experimental Texts by TOEIC score group

Examining the mean reading speed for each text type within the respective reading ability class found that the mean reading speed of MT-EJ and MT-J texts was significantly faster than that of ENG in 400-class; the mean speed of MT-J was significantly faster than that of ENG in 500-class, 600-class, 700-class, and 800-class; none of the four experimental texts were read significantly faster than ENG in 900-class.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	–	↑	↑	–
500-	–	↑	–	–
600-	–	↑	–	–
700-	–	↑	–	–
800-	–	↑	–	–
900-	–	–	–	–

Table 10: The Reading Speed (ENG)

Further examination revealed that the mean speed of JPN is significantly faster than the speed of any other text types.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	↓	↓	↓	↓
500-	↓	↓	↓	↓
600-	↓	↓	↓	↓
700-	↓	↓	↓	↓
800-	↓	↓	↓	↓
900-	↓	↓	↓	↓

Table 11: The Reading Speed (JPN)

5.4 Reading Efficiency

We examined effective reading speed (eWPM), which was derived from reading speed divided by comprehension test score, and compared the effciency between the controlled texts, i.e., ENG and JPN, and the experimental texts. The mean eWPM of controlled texts was marked 50.63 eWPM (S.D. 32.30) in ENG texts and 135.664 (S.D. 72.919) in JPN texts. The ENG texts were read the most slowly, and the JPN texts were read the fastest.

Features	Mean Score	S.D.
CHU	51.18	36.55
MT-J	71.60	44.61
MT-EJ	53.66	30.91
MT-W	47.39	34.41
ENG(Control)	50.63	32.30
JPN(Control)	135.66	72.92

Table 12: The Mean Reading Speed

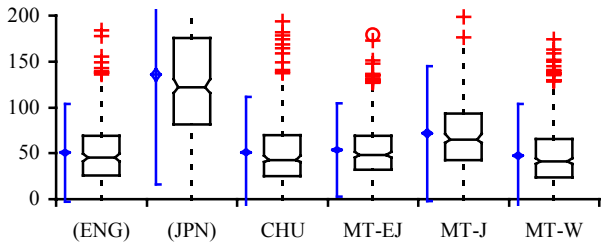

Figure 4: The Mean Effective reading Speed

The following tables show the mean effective reading speed for each text types. Table 13 presents the mean efficient speed of control texts, and Table 14 presents the mean efficient speed of experimental texts.

The mean efficienct speed of ENG is consistent with the ranking of TOEIC score. By contrast, the mean score of JPN seems inconsistent with the ranking.

Class	Text Type	
	ENG	JPN
400-	34.95	131.20
500-	33.93	143.80
600-	50.17	122.42
700-	52.72	151.38
800-	60.08	134.68
900-	77.94	134.77

Table 13: The Mean Efficient Speed of Controlled Texts

The following table shows the mean speed of each text type for each TOEIC score group. If the increase of efficient speed exceed 5 point compared with the weighted speed of ENG, an up-arrow appears. When the weighted speed decreases in more than 5 point, a downward arrow comes up. The crossbar indicates that the weighted speed change does not exceed 5 point.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	32.09 –	75.01 ↑	54.77 ↑	34.76 –
500-	38.98 ↑	70.33 ↑	44.24 ↑	35.65 –
600-	44.78 ↓	67.39 ↑	45.61 –	35.53 ↓
700-	52.83 –	68.66 ↑	54.99 –	31.53 ↓
800-	73.13	82.20 ↑	66.44 ↑	70.81 ↑
900-	73.68 –	71.12 ↓	66.75 ↓	69.55 ↓

Table 14: The Mean Speed of Experimental Texts

Examining the mean effective reading speed for each text type within the respective reading ability class found that the mean effective reading speed of MT-J texts was significantly greater than that of ENG in 400-class and 500-class; the mean speed of MT-J and MT-W was significantly faster than that of ENG in 600-class; none of the four experimental texts were read significantly faster than ENG in 700-class, 800-class and 900-class.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	–	↑	–	–
500-	–	↑	–	–
600-	–	↑	–	↑
700-	–	–	–	–
800-	–	–	–	–
900-	–	–	–	–

Table 15: The Reading Speed (ENG)

Further examination revealed that the mean efficient speed of JPN is significantly greater than the speed of any other text types.

Class	Text Type			
	CHU	MT-J	MT-EJ	MT-W
400-	↓	↓	↓	↓
500-	↓	↓	↓	↓
600-	↓	↓	↓	↓
700-	↓	↓	↓	↓
800-	↓	↓	↓	↓
900-	↓	↓	↓	↓

Table 16: The Reading Speed (JPN)

6 Conclusion

In this paper, we contended that reading support systems should be evaluated not only with comprehension performance but also with speed performance. In our experiment, we confirmed that the reading speed-based evaluation exhibited the similar effect to the comprehension-based evaluation as the previous studies suggested. The experimental result presented that the efficiency of the systems varied, depending on reading ability of users, as the previous study (Ohguro 1993; Fuji et al. 2001) confirmed.

We have not yet examined the effect of supporting systems at the sentence-level, but we will expect the further study would reveal it.

References

J. C. Alderson. 2000. *Assessing Reading*. Cambridge University Press, Cambridge.

R. Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology* 32: 221-233.

M. Fuji, N. Hatanaka, E. Ito, S. Kamei, H. Kumai, T. Sukehiro, T. Yoshimi and H. Isahara. 2001. Evaluation Method for Determining Groups of Users: Who Find MT "Useful." *Proceedings of the MT Summit VIII*.

P. Muter and P. Marutto. 1991. Reading and Skimming from Computer Screens and Books: The Paperless Office Revisited? *Behaviour & Information Technology* 10:257-266.

Y. Ohguro. 1993. Evaluating the Validity of Printing Japanese Words alongside English Text. *Technical Report on Information Processing Society of Japan*, 93-NL-79: 127-134.

K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics*: 311-318.

T. Yoshimi, K. Kotani, T. Kutsumi, I. Sata and H. Isahara. 2005. A Method of Measuring Reading Time for Assessing EFL-Learners' Reading Ability. *JSiSE* 22: 24-29.