IEMS Vol. 3, No. 1, pp. 85-91, April 2004.

Relative Difficulty of Various English Writings by Fuzzy Reasoning and Its Application to Selecting Teaching Materials

Hiromi Ban†
Faculty of Human Sociology
Toyama University of International Studies, Oyama-machi, Toyama, 930-1292, JAPAN
Tel: +81-76-483-8000, Fax: +81-76-483-8008, E-mail: je9xvp@yahoo.co.jp

Toby Dederick
Faculty of Foreign Languages
Hokuriku University, Kanazawa-shi, Ishikawa, 920-1180, JAPAN
Tel: +81-76-229-1161, Fax: +81-76-229-1393, E-mail: dederick@nsknet.or.jp

Hidetaka Nambo
Faculty of Technology
Kanazawa University, Kanazawa-shi, Ishikawa, 920-8667, JAPAN
Tel: +81-76-234-4835, Fax: +81-76-234-4900, E-mail: nambo@ec.t.kanazawa-u.ac.jp

Takashi Oyabu
Faculty of Economics
Kanazawa Seiryo University, Kanazawa-shi, Ishikawa, 920-0813, JAPAN
Tel: +81-76-253-3962, Fax: +81-76-253-3995, E-mail: oyabu@seiryo-u.ac.jp

Abstract. The writing styles of *TIME* and *Newsweek* are analyzed using a specially developed linguistic program. These two news magazines were chosen because of their wide popularity. As for the results, it became obvious that both the frequency curve of words and that of characters have not changed for the past 60 years. Also, we have found that the frequency curves have some inflection points and that the genre of English writings can be identified by these points. After counting the percentage of required vocabulary for junior high school students and high school students in English writings, we can derive the relative difficulties of them using fuzzy reasoning. Fuzzy rules are constructed using features of the characteristic curves. We feel it would be a good guide index when selecting textbooks or supplementary readers.

Keywords: English style analysis, English education, teaching materials analysis, fuzzy reasoning

1. INTRODUCTION

In the classrooms of English language, various textbooks and supplementary readers are in use. Teachers must recognize their levels of difficulty in order to attain their objectives efficiently. This paper attempts to examine, by drawing the frequency curves of words and characters from several kinds of English writings, what levels of difficulty they actually have as textbooks. The materials we chose to analyze are: *TIME* (Material 1); *Computing Essentials*, a technological writing for general people by Don Cassel, 1994 (Material 2); an essay by Mike Royko called *A Selection of 2O Columns from DR. KOOKIE, YOU'RE RIGHT!*, 1989 (Material 3); and three literary works: *The Bridges of Madison County* by Robert James Waller, 1992 (Material 4), *The Old Man and the Sea* by Ernest Hemingway, 1952 (Material 5) and *Sarah, Plain and Tall* by Patricia MacLachlan, 1985 (Material 6).

After scanning these writings and transferring them into digital data by an OCR system, we analyzed them using a program we specially prepared in C++ (Ban *et al.*, 1999). To attain the degree of relative difficulties, the program counted the percentage of required vocabulary for Japanese junior and senior high school students, then applied simple fuzzy rules for reasoning to analyze the outcome.

†: Corresponding Author

As a result, it became clear that any material we chose is too difficult for Japanese junior high school students, that Material 1 is very difficult even for senior high school students, and that Material 5 and Material 6 have the same degree of difficulty as standard textbooks widely used in senior high schools in Japan. Though this method of analysis is experimental, it could be employed as an index guide in selecting textbooks or supplementary readers.

2. THE CHARACTERISTICS OF EACH GENRE OF ENGLISH WRITINGS

First, we analyzed the frequency of words used in the past 50 years of *TIME*, the most popular news magazine in the U.S. The results have already been reported previously (Ban *et al.*, 1998). As examples, the results for *TIME '90* and *'97* are shown in Figure 1 and Figure2 respectively. These are the analysis of the first issues in 1990 and 1997. The vertical axis stands for word frequency, the horizontal axis for the order of frequency. Moreover, we can claim the same result came from the analysis of the first issue of *Newsweek* in 1997 (Figure 3). While a slight fall of the curve is observed around the 13th order, and the frequency in the latter orders seems comparatively high, the same tendency from the 1940s to the 1990s is observed as a whole. Furthermore, when comparing this result with that of *TIME*, little difference is observed as a whole.

As to the frequency of characters in *TIME*, Figure 4 and Figure 5 show the results of *TIME '90* and *'97* respectively. Although some changes in characters occur, and in the case of *TIME '97*, the frequency is extremely low in the 49th and 50th, these curves are little changed from the previous reports as whole. Figure 6 shows the frequency of characters in *Newsweek*. Similarly, it can be said in this case that the total tendency is little changed, while in the lower half the curve differs a little with the years. When also comparing this result with that of *TIME*, little difference is observed at least in the upper half.

From the analysis of *TIME* and *Newsweek* on the whole, it becomes evident that the frequency curves of words and characters have not changed for the past 60 years. It is also possible to analyze a variety of English writings by comparing the other materials mentioned above with these curves.

For example, the characteristics of the frequency of words in *The Old Man and the Sea* (Material 5), which has the same

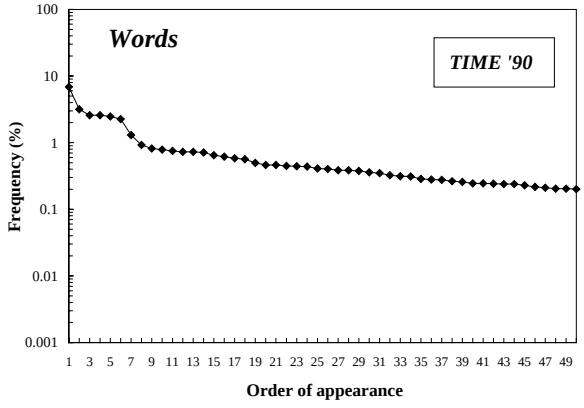

Figure 1. Frequency characteristics of word-appearance for *TIME '90*

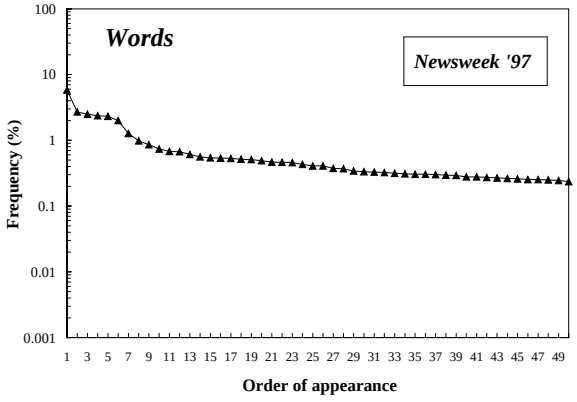

Figure 3. Frequency characteristics of word-appearance for *Newsweek '97*

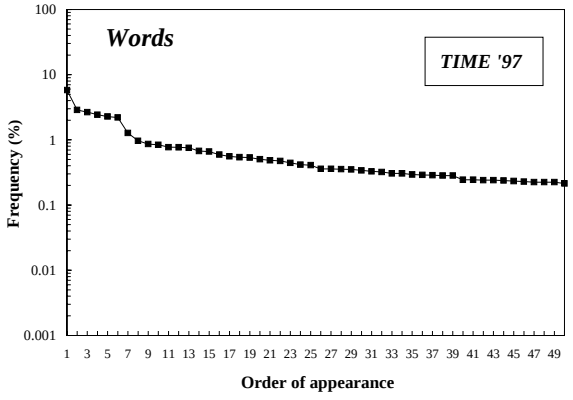

Figure 2. Frequency characteristics of word-appearance for *TIME '97*

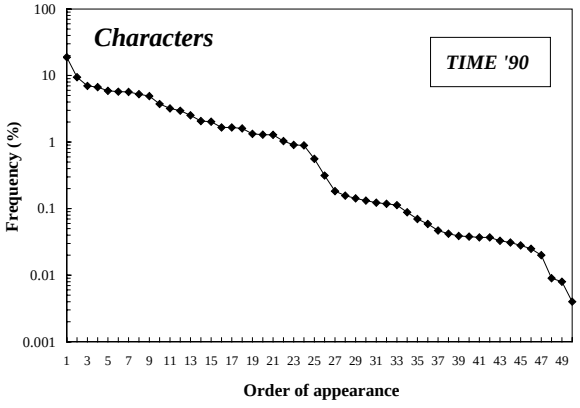

Figure 4. Frequency characteristics of character-appearance for. *TIME '90*

level of difficulty as supplementary readers used in senior high schools, proves different from that of *TIME* (Figure 7). That is, while the curve of Material 5, in the early stage, is formed a little above that of *TIME*, it becomes lower at the 4th frequency, and higher again at the 7th. As for the frequency of characters, it suddenly decreases from around the 30th (Figure 8), because words and characters used in the textbooks adopted in Japanese high schools are limited. On the other hand, as journalistic writings contain more extensive materials, the latter half of the frequency curve tends to decline more gently.

As a result of examining the frequency curve of each genre, there proves to be an inflection point where the value suddenly declines. Table 1 shows the order in which each inflection point appears. If the characteristics of the genre examined are close to a journalistic one, the order corresponds to the words 'in' and 'that'. Journalistic writings often use demonstrative pronouns and expressions about place or location. That is the reason why the frequent use of 'in' and 'that' is to be observed in them. In journalistic writings, relative pronouns seldom appear, while demonstrative pronouns are often in use instead. On the other hand, in literary works, an inflection point occurs at the order that contains personal pronouns such as 'I' and 'he', according to the narrative each literary work adopts. In literary works, it is frequently observed that the inflection point occurs at a higher order than 6.

Furthermore, we compared these materials at 9 linguistic points and drew a radar chart so that features of each of these points might be judged visually (Fig. 9). The standard of the chart is based on *TIME '97*. Whereas the charts of *The Bridges of Madison County* (Material 4) and

A Selection of 20 Columns form DR. KOOKIE, YOU'RE RIGHT! (Material 2) are shaped crooked compared to *TIME*, that of *Computing Essentials* looks well-proportioned to *TIME*. The reason may be that literary works are more affected by the author's manners of writing than that of technical writings.

3. PERCENTAGE OF REQUIRED VOCABULARY FOR JUNIOR AND SENIOR HIGH SCHOOLS IN THE MATERIALS

Next, we examined the materials in terms of percentage of required English vocabulary for Japanese junior and senior high

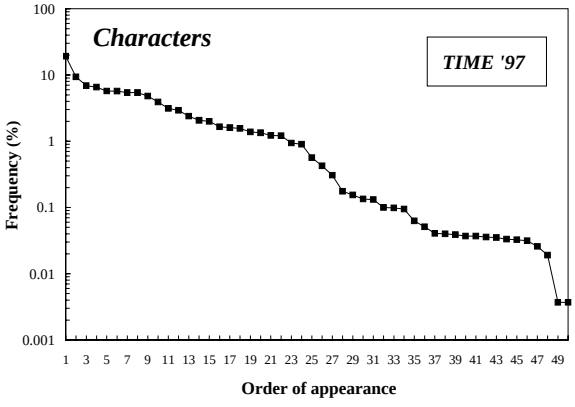

Figure 5. Frequency characteristics of character-appearance for *TIME '97*

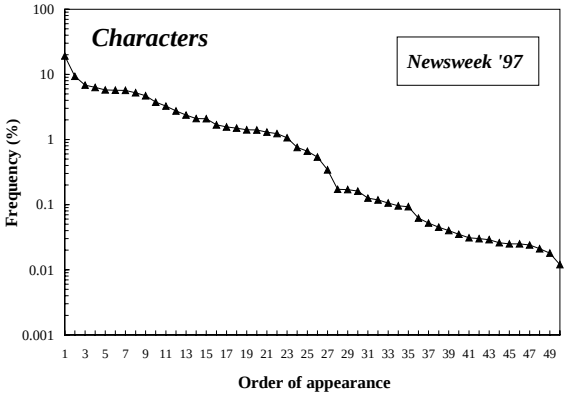

Figure 6. Frequency characteristics of character-appearance for *Newsweek '97*

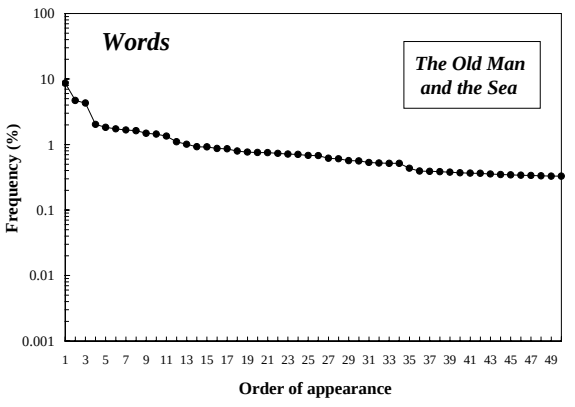

Figure 7. Frequency characteristics of word-appearance for *The Old man and the Sea*.

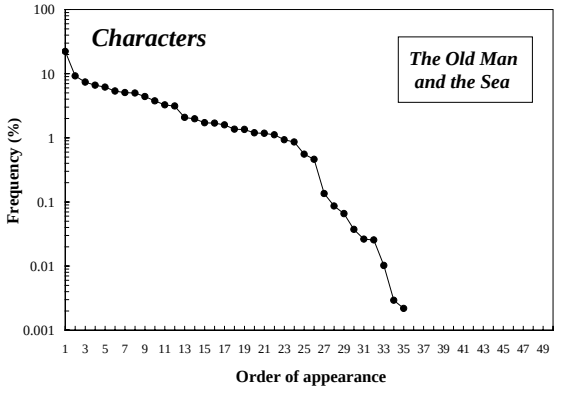

Figure 8. Frequency characteristics of character-appearance for *The Old Man and the Sea*.

Table 1. Order and vocabulary where inflection point occurs on frequency curve

	Genre	Materials	Inflection point Order	Inflection point Vocabulary
Tecchnical Writings & Journalism ↑	Technical Writings	*Computing Essentials*	6	AND
			7	IN
			8	THAT
	Journalism	*TIME*	6	IN
			7	THAT
		Newsweek	6	IN
			7	THAT
	Columns	*A Selection of 20 Columns from DR. KOOKIE, YOU'RE RIGHT!*	5	I
			6	OF
	Textbooks (H.S.)	*MILESTONE English Reading*	5	A
			6	IN
	Textbooks (J.H.S.)	*SUNSHINE ENGLISH COURSE 3*	3	A
			4	OF
	Literature (J.H.S.)	*Sarah, Plain and Tall*	2	AND
			3	I
↓ Literature	Literature (H.S.)	*The Old Man and the Sea*	3	HE
			4	OF

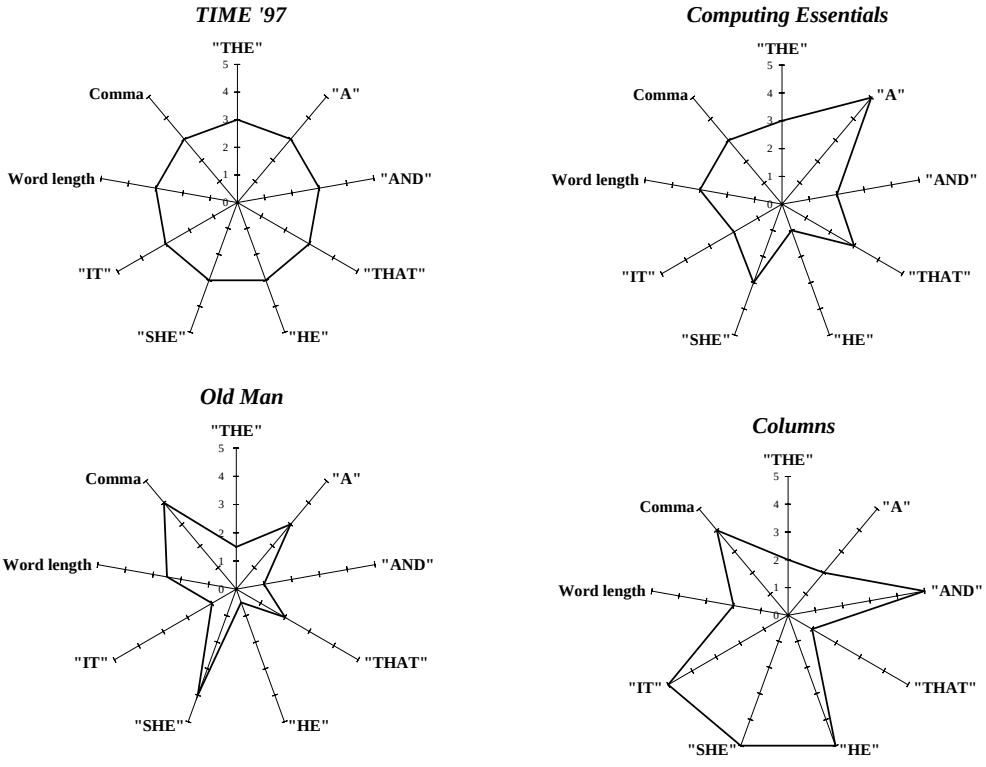

Figure 9. Rader charts showing characteristics of each material

schools using four criteria. We used the words from the required vocabulary for junior high school students selected by the Ministry of Education (507 words), "the words that appeared in more than 5 publishers' textbooks out of 7" presented in *English Words in the Textbooks of Junior High School Students* (ed. Fumio Akao, Obunsha, 1995), hereafter, called 'important words for junior high school students' (233 words), and the most important words (550 words) and important basic words (1600 words) for senior high school students selected in *Basic 3800 English Words: for Entrance Examination of University* (ed. Yoshio Akao, Obunsha, 1997).

The percentage of these words in each material is shown in Table 2. To take the example of *TIME*, the percentage of frequency of required vocabulary for junior high school, in

Table 2. Proportion of junior and senior high school vocabulary in each material

		Word frequency (%)				Word type (%)			
		J.H.S. Required	J.H.S. Important	H.S. Most important	H.S. Important	J.H.S. Required	J.H.S. Important	H.S. Most important	H.S. Important
TIME '90		51.4	6.5	5.4	10.1	8.9	3.1	5.7	18.2
Computing Essentials		55.1	4.4	6.4	13.2	16.8	4.7	11.9	23.8
(Literature)	*Madison*	63.4	10.0	3.8	7.3	15.1	6.3	7.7	21.8
	Old Man	71.2	9.3	4.3	5.7	22.3	6.9	8.2	20.5
	Sarah	64.1	9.0	2.2	4.5	33.2	7.9	5.8	14.7
Columns		63.4	8.2	4.8	7.3	17.2	6.3	9.4	22.1
Textbooks (J.H.S.)	*SUNSHINE 1*	76.7	13.2	0.6	1.4	66.2	13.2	1.9	3.5
	SUNSHINE 2	72.3	13.7	1.2	2.6	51.7	16.7	3.0	6.9
	SUNSHINE 3	71.8	12.5	3.4	3.7	47.7	15.8	8.5	8.6
Textbooks (H.S.)	*MILESTONE 1*	67.1	10.8	4.4	5.9	29.7	11.1	10.1	18.4
	MILESTONE 2	65.8	10.3	5.2	7.9	26.3	9.5	11.2	22.4
	MILESTONE Reading	65.8	9.4	5.4	7.5	20.9	7.4	10.6	24.6

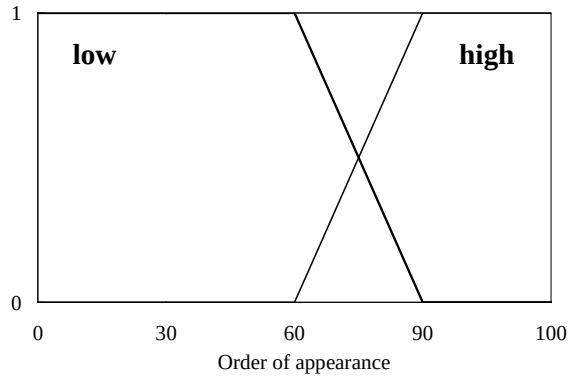

Figure 10. Membership function of word frequency

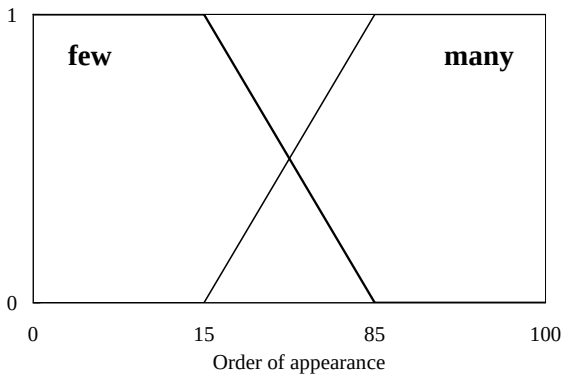

Figure 11. Membership function of word type

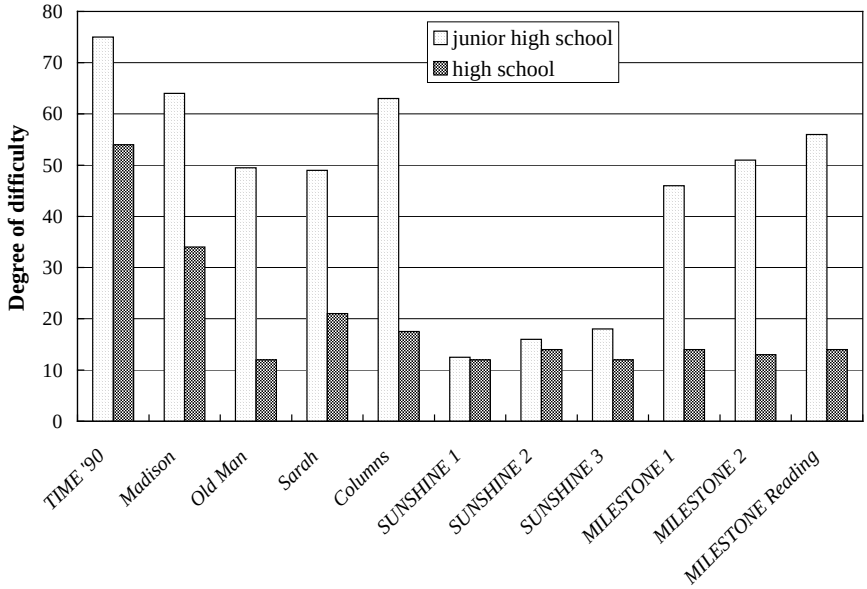

Figure 12. Degree of difficulty estimated by fuzzy reasoning

terms of appearance, is around 50%. As to the important words for junior high school, less than 60% appear. Considering even the important senior high school words, less than 70% appear. On the other hand, the textbook for junior high school

(*SUNSHINE*, Kairyudo) for the first-year students includes about 90% of the important words for junior high school, for the second-year students about 86%, and for the third-year students about 84%. The higher the grade is, the slightly the lower the percentage. In addition, the textbook for senior high school (*MILESTONE English*, Keirinkan) includes 75%.

We also looked at the frequency of the words in terms of type. For example, *TIME* contains only 12% of the words contained in both required and important words for junior high school students. Even if we count the most important and important basic words for senior high school students, the result is about 36%. Therefore, *TIME* is a very difficult material for even senior high school students.

The Old Man and the Sea (Material 5) contains 71% of the required vocabulary for junior high school students in terms of appearance, but in terms of type, there are less than 30% of even the important words for junior high school students. Therefore Material 5 is very difficult for junior high school students. But if we count the words for senior high school students, the percentage is considerably raised up to 58%. Therefore it may be said that Material 5 is an efficient supplementary reader for senior high school students. Another literary work, *Sarah, Plain and Tall* (Material 6), in terms of type of vocabulary, contains 64% of the words for senior high school students, but in terms of appearance, about 80%, slightly lower than Material 5 (90%). Therefore Material 6 is somewhat more difficult than Material 5.

From all of the above mentioned, it may be said that if we analyze the required or important words for junior and senior high school students, finding the frequency curve of each material, and calculating the percentage of the words contained in it, then the degree of relative difficulty of the material can be roughly estimated. But to estimate it more precisely, the rules by which we actually judge the difficulty of textbooks are to be applied to this process. This study adopted a set of fuzzy rules and fuzzy reasoning because human sensitivity about difficulty is vague and ambiguous.

4. ESTIMATING DIFFICULTY BY FUZZY REASONING

We defined the following 4 rules to estimate the difficulty of the material by the frequency of appearance of the words and the frequency of the type of words. Because this study is a preliminary one which aims to estimate difficulty by fuzzy reasoning, the rules are limited to that end and to the most basic ones. To satisfy the needs of actual classrooms, more diverse and complex rules would be required.

Rule 1: If both the frequency of appearance and the frequency of type are high, then the degree of difficulty is low.
Rule 2: If the frequency of appearance is low and the frequency of type is high, then the degree of difficulty is average.
Rule 3: If the frequency of appearance is high and the frequency of type is low, then the degree of difficulty is average.
Rule 4: If both the frequency of appearance and the frequency of type are low, then the degree of difficulty is high.

The membership functions corresponding to the frequency of appearance and the frequency of type are defined in Figure 10 and Figure 11 respectively. Figure 12 shows the degree of difficulty estimated by this reasoning. The values that are lightly dotted show the degree of difficulty resulting from the sum of the required and important words for junior high school students. It shows that *TIME* has a degree of difficulty of 75%, and the difficulty is 5 times more than the textbook for junior high school students (*SUNSHINE*). Literary works (Materials 1, 2, 3) mark at 50% to 60%. Especially, *The Bridges of Madison County* (Material 4) turned out to be difficult. The degree of difficulty it has is more than 2 times of that of the textbook for senior high school (*MILESTONE*). *The Old Man and the Sea* (Material 5) is suited for the level of senior high school students, but to junior high school students, it must be considerably difficult material.

The degree of difficulty for senior high school students, as is the case for junior high school students, is estimated from the frequency of appearance and type of the sum of the most important and important basic words for senior high school students. The result of this shows that the textbooks for junior and senior high school students show a similar degree of difficulty for senior high school students. One of the reasons for this may be that the reasoning is based only on words, not on idioms, phrases, and structures of sentences.

5. CONCLUSIONS

We have tried to estimate the degree of relative difficulty of each genre of English writings from the frequency of appearance and type of English words used in textbooks for Japanese junior and senior high school students. This paper is an experimental one which aims to seek a clue to estimate the degree of difficulty by applying fuzzy rules to the frequency data.

As a result, the estimation which teachers in classrooms approximately would make has been acquired. Also, at what times each material is more difficult than the textbook adopted in each school can be estimated from this reasoning. *The Old Man and the Sea*, which is widely adopted as a supplementary reader in senior high schools, has turned out to be as difficult as textbooks for senior high school students, while *TIME* has turned out to be 4 times more difficult, and *The Bridges of Madison County* nearly 3 times more difficult than ordinary textbooks. What is required in the future is to make a more practical estimation by forming more rules for fuzzy reasoning.

REFERENCES

Ban, H., Oyabu, T., Sugata, T. and Dederick, T. (1999) Statistical Characteristics of Prepositions in English Newspapers of Japan,

the United States and the United Kingdom. *Proceedings of the 3rd International Conference on Engineering Design and Automation*, Vancouver, Canada, 216-221.

Ban, H., Sugata, T., Dederick, T. and Oyabu, T. (1998) Metrical Characteristics of English Writings Using an Exponential Function. *Proceedings of the 1st Korea-Japan Joint Conference on Industrial Engineering and Management*, Taejon, Korea, 47-50.