Towards building a usable corpus collection for the ELT classroom

Kiyomi Chujo, Masao Utiyama and Chikako Nishigaki

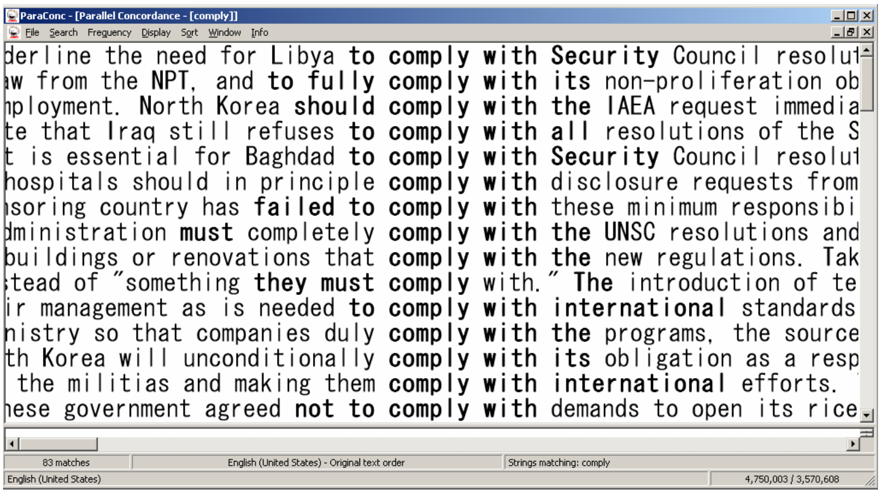
Nihon University, National Institute of Information and Communications Technology, and Chiba University

Abstract

As tantalizing as the potential for corpus application is in second language acquisition, we educators seem to stumble over how to make concordancing lines understandable for learners. This study explores various criteria for evaluating text samples by difficulty level in order to provide a collection of rated parallel English-Japanese corpus texts which educators can use in ELT classrooms, and provides the rating tools and methodology so that educators can evaluate their own classroom material. Data was collected from two English-Japanese parallel corpora, and seven indices (readability scores, average word length, Japanese school textbook vocabulary coverage, BNC text coverage, Japanese vocabulary ratio, sentence length, and kanji character ratios) were applied to measure the linguistic difficulty of both English and Japanese text samples. It was noted that most of the texts in the collection were advanced level, and that there is a shortage of copyright available e-text data at the beginner level. Nevertheless, this study identifies several applicable indices, provides a rated collection of titles at varying levels of difficulty, and takes corpus usage one step closer to its ideal application.

1. Introduction

Technology is changing our world and provides us with new tools for learning. Using computers to produce corpora and concordancing data provides us with exciting new possibilities in our daily language-learning environment. Although recognised by educators as a potentially useful tool, corpus application has both highly contested advantages and disadvantages. Few attempts have been made to use corpora directly in the classroom by foreign language teachers and learners in Japan other than students of linguistics because of the difficulty students have understanding the concordance examples retrieved (Tono 2003). One English as a Foreign Language (EFL) learner looking at Figure 1 might easily be overwhelmed by not only the long list of examples, but by vocabulary too advanced to be useful. Addressing this widely acknowledged barrier to corpus application is the subject of this paper. How can we as educators simplify concordancing lines to make them understandable, and therefore useful, to learners? Aston's advice (2001: 43) is to carefully select the corpus or subcorpus; and Thomas (2002) has provided a discussion of post-concordance filtering according to each word frequency; however, to date no one has developed an objective, easy-to-use criterion for evaluating the linguistic difficulty of various texts; and little research, if any, has tried to specifically investigate in what way or ways the corpora texts are difficult for foreign language learners.



```
derline the need for Libya to comply with Security Council resolut
aw from the NPT, and to fully comply with its non-proliferation ob
nployment. North Korea should comply with the IAEA request immedia
te that Iraq still refuses to comply with all resolutions of the S
t is essential for Baghdad to comply with Security Council resolut
hospitals should in principle comply with disclosure requests from
nsoring country has failed to comply with these minimum responsibi
dministration must completely comply with the UNSC resolutions and
buildings or renovations that comply with the new regulations. Tak
stead of "something they must comply with." The introduction of te
ir management as is needed to comply with international standards
nistry so that companies duly comply with the programs, the source
th Korea will unconditionally comply with its obligation as a resp
 the militias and making them comply with international efforts.
nese government agreed not to comply with demands to open its rice
```

83 matches | English (United States) - Original text order | Strings matching: comply

English (United States) | 4,750,003 / 3,570,608

Figure 1. *Comply* in a monolingual corpus

2. Research goals

To have understandable concordancing lines, we must begin with an understandable corpus. As Aston (2001) points out, the texts chosen for the corpus must be selected carefully at an appropriate level for the learner. Therefore the goals of this study were twofold: (1) to identify effective, easy-to-use criteria for evaluating the linguistic difficulty of various English and Japanese texts or subcorpora; and (2), once identified, to measure the linguistic difficulty of the various English and Japanese texts. This was done by creating a large parallel corpus, extracting text samples, and then measuring the text samples with several indices. The end product of this research is a collection of level-defined subcorpora to be used for direct classroom application as well as the tools for educators to apply to their own texts. The criteria can be used for selecting appropriate-level data-driven learning material and reading textbooks, and in assisting web-searchers in choosing appropriate-level webpages. Since the level of Japanese in the parallel corpus data was also evaluated, it is hoped that this material is useful for learners of Japanese as well as English learners.

3. Method

To create a main corpus, a large number of text samples were collected into one Parallel Corpora Text Collection (hereafter 'the Collection'). The text samples were extracted from the following two parallel corpora: (1) the English-Japanese

Translation Alignment Data,[1] which has 84 narrative and expository texts or subcorpora written originally in English (632,564 words) and translated into Japanese (1,011,873 morphemes), and aligned manually; and (2) the Japanese-English News Article Alignment Data,[2] which comprises 180,000 sentence pairs from *The Yomiuri Shimbun* (6.1 million Japanese morphemes) and *The Daily Yomiuri* (4.9 million English words), automatically aligned. This resulting Collection therefore has two types of texts: stories and documents, and newspaper articles. The story/document division contains 63 titles, encompassing a wide selection of texts including stories (e.g. *Jack and the beanstalk*, *The black cat*), reading material in content areas (e.g. *The Darwinian hypothesis, The declaration of independence*) and blogs (e.g. *Freedom or copyright?*). All titles included in this Collection's story/document division were readily available e-texts either with granted reproduction and redistribution rights by the copyright holders, or already in the public domain. Unfortunately, the availability of these titles is limited, most notably at the beginner level.

4. Referential data

In order for this study to be meaningful in an EFL or JFL (Japanese as a Foreign Language) context, we must compare the vocabulary of the Collection to a standard.[3] In this case, a comparison of the English text vocabulary was made to the vocabulary learned by Japanese students by calculating text coverage with the top selling series of junior and senior high school (hereafter JSH) textbooks in Japan from the 7th through 12th grades.[4] This vocabulary, totalling 3,098 different words, is representative of the vocabulary studied by most college students before entering university. The high-frequency words from the British National Corpus (BNC) were also explored as criteria. The certified standards of the *Japanese Language Proficiency Test: Test Content Specifications* (hereafter *JLPT Test Content Specifications*, see Kokusai Kouryuu Kikin 2002) served as a reference guide for the Japanese texts.

5. Text samples

A total of 99 sets of both English and Japanese sample texts were extracted from each of the two corpora discussed above. From the first – story and document – corpus, 87 sets of both English samples (on average 2,076 words) and Japanese samples (on average 2,981 morphemes) were randomly extracted from the 63 titles. When the original text was small, the entire text was selected as a sample. When the whole text was larger than the capacity of the Japanese language analysis programme,[5] two sets of samples were randomly extracted. English and Japanese samples are shown in Figure 2. From the second – newspaper – corpus, 12 sets of both English samples (on average 2,294 words) and Japanese samples (on average 2,963 morphemes) were selected randomly.

English text

Every afternoon, as they were coming from school, the children used to go and play in the Giant’s garden. It was a large lovely garden, with soft green grass. Here and there over the grass stood beautiful flowers like stars, and there were twelve peach trees that in the spring time broke out into delicate blossoms of pink and pearl, and in the autumn bore rich fruit. The birds sat on the trees and sang so sweetly that the children used to stop their games in order to listen to them. “How happy we are here!” they cried to each other. One day the Giant came back. He had been to visit his friend the Cornish ogre, and had stayed with him for seven years. After the seven years were over he had said all that he had to say, for his conversation was limited, and he determined to return to his own castle. When he arrived he saw the children playing in the garden.

from *The Selfish Giant*, Oscar Wilde

Japanese text

子どもたちは毎日、午後になって学校から帰ってくると、大男の庭に行って遊ぶのが常でした。そこは、柔らかい緑の草が生えた、広くて素敵な庭でした。草むらのあちこちには、星に似た美しい花が立っておりました。その庭には十二本の桃の木があり、春になると薄桃色と真珠色の繊細な花があふれるように咲き、秋には豊かな果実が実ります。鳥たちは木々の上でたいそう甘い歌声を響かせるので、子どもたちは遊ぶのをやめて聞きいるのでした。「ここで遊ぶのはなんて楽しいんだろう！」と、くちぐちに声をあげました。ある日、大男が帰ってきました。彼はコーンウォールに住む鬼の友人を訪問し、そこで７年間いっしょに過ごしていました。７年が過ぎ、話したいことは全部話したし、もう話題もなくなってきたので、自分の城に帰ろうと思った

Figure 2. Text samples

6. Indices investigated

Before deciding on the specific indices to apply in this study, the educational literature was examined to understand which indices had been applied to measure linguistic difficulty, and of those, which might be the most useful for this study. We identified seven: four for English texts, and three for Japanese texts. The indices were applied using various computer programmes; in cases where the whole text was too large for the software programme used to measure the index, two sets of samples were extracted from one title, and the average score of these two samples was used as the title’s representing difficulty score. For the English texts, the indices included (1) readability scores, (2) average word length, (3) the text coverage of JSH textbook vocabulary, and (4) text coverage from the BNC. We chose indices which could be applied by using readily available software so that teachers would be able to apply these indices to their own data without having to develop programmes themselves.

For JFL learners, it is Japanese that is the focus for learning; therefore it was also important to measure the text difficulty of the Japanese text samples.

JLPT Test Content Specifications state that the Japanese text content itself cannot be measured for difficulty, but that it is possible to measure indices for quantitatively controllable variables on text difficulty such as vocabulary, sentence length and the number of kanji characters (Kokusai Kouryuu Kikin 2002: 219). Therefore to evaluate the Japanese samples, the indices applied were (5) the *Test Content Specifications* Levels 1 and 2 vocabulary ratios, (6) sentence length, and (7) the percentage of kanji (characters). Each step is outlined below.

7. English readability

7.1 Readability formulas

The term 'readability' refers to the factors that affect understanding a text and therefore success in reading. In the context of this study, readability particularly includes the complexity of words and sentences in relation to the reading ability of the reader. We use the term 'reading grade level' to indicate the reading grade of a text that could be read and just understood by a student of that grade who has average reading ability; for example, a score of 8.0 means that an average native-speaking eighth grader would understand the text. Objective measures of readability are generally done either by comparing a text with a standard word list or utilising calculations involving the sentence length and number of syllables.

For this study, we calculated the readability score by using Readability Calculations software,[6] which contains nine widely used formulas including the Flesch-Kincaid formula (Flesch 1974). In our preceding study, we applied these formulas to more than 100 text samples of various genres to observe the difference among the yielded scores by different formulas (Chujo *et al.* 2004). It was noted that three formulas, e.g. the Flesch-Kincaid formula, the SMOG formula (McLaughlin 1969) and the Fry Graph (Fry 1968), were the most reliable, as demonstrated by the range of grades observed from the difference between each formula's maximum and minimum grade level, and also from the appropriate within-group variability by the standard deviation. In order to provide more validity for the current study and in an attempt to calibrate some fixed points on the scale of readability, the averages of those three formulas were applied to the samples and the results are used to express 'readability'.[7]

7.2 Comparing Japanese and English readability scores

Since much of the work on readability formulas has been done in the US, the formulas give a numerical value representing an American grade level. In order for these to be relevant to Japan's educational situation, the readability of a representative sample of each JSH textbook by grade level was also calculated to provide comparable measures to the readability scores of the targeted samples. The procedure was as follows:

(1) From each of the junior high school textbook series *Horizon 1*, *2* and *3*, which corresponds to the US 7th (Japan: junior high 1st), 8th (junior high 2nd) and 9th (junior high 3rd) grades respectively, two reading lessons entitled 'Let's read' were selected, giving a total of six samples.

(2) In the Japanese senior high school textbook series *Unicorn I*, *II* and *Reading*, *Unicorn I* corresponds to the US 10th grade (Japan: senior high 1st), *Unicorn II* corresponds to the 11th (senior high 2nd), and *Unicorn Reading* corresponds to the 12th (senior high 3rd). From each of the textbooks, three lessons (Lessons 1, 5 and 10) were selected, giving a total of nine samples.

(3) The readability scores of these 15 samples from the Japanese JSH English textbooks in terms of American grade levels were measured by using the same readability measures (the average of the three readability formulas) as a scale.

Figure 3 shows the reading grade levels of the JSH textbooks investigated. The vertical bars on the graph indicate the reading grade levels predicted by the use of the three readability formulas. For example, the readability of the Japanese first year junior high school textbook (US 7th grade) was rated as a (US) 2.8 reading grade. In other words, the English contained in the Japanese first year junior high textbook might be readable by a US second grader nearing the end of the school year. The reader may recall that two samples at the junior high (JH) level and three samples at the senior high (SH) level were used; these are averaged together in the graph below.

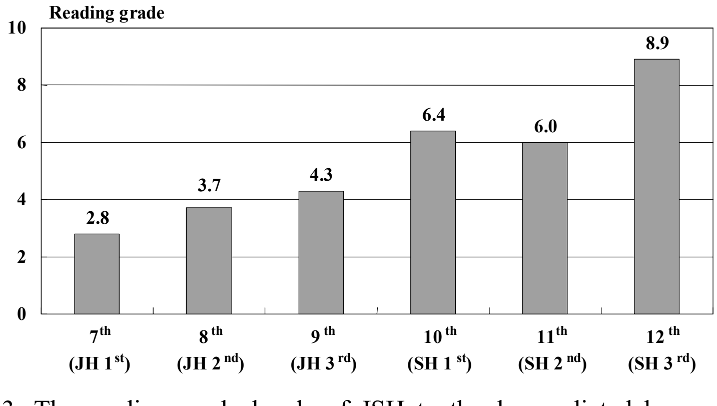


Figure 3. The reading grade levels of JSH textbooks predicted by readability formulas

Looking at Figure 3, we can see that the graduation of reading grade levels among each grade textbook appears almost as one might expect. The 7th grade JH 1st textbook ranks the lowest, followed by 8th and 9th grade texts, then the grade level rises sharply to the 10th, then decreases slightly at the 11th, and finally reaching the highest level with the 12th (SH 3rd) textbook. This graph indicates that the readability transition occurring in textbooks from JH to SH is neither smooth nor easy. A discussion of the validity of the readability in Japanese English textbooks is an interesting and necessary discussion but one which goes beyond the scope of the present article; nevertheless, the data in Figure 3 does provide a means to compare measures for the readability scores of the target samples. For the purposes of this study, we created three categories of linguistic difficulty based on these results: texts lower than the 5.9 grade level were termed 'Level I,' those falling between the 6.0 and 8.9 grade levels are 'Level II', and those higher than the 9.0 grade level are 'Level III'.

7.3 Average word length

Using Writers' Workbench Version 8.15 Style Statistics,[8] we obtained information on basic stylistic variables such as average word length, sentence length, the number of simple and complex sentences, the percentage of *to be* verbs compared to the total number of verbs, and the percentage of use of passive voice. Of these variables, only average word length provided applicable information for this project. We can speculate that the other indices did not work well because the sample text sizes may have been too small. Also note that none of the three readability formulas described above uses 'average word length' by calculating the number of letters, so measuring average word length provides separate and additional data. As an added advantage, this index is easy to use and intuitively easy to understand.

7.4 JSH text coverage

The next calculation was the extent to which the vocabulary in the JSH texts does or does not cover the vocabulary used in each of the text samples.[9] This constitutes one way of obtaining an accurate estimate of the vocabulary level of each text, which is crucial information to EFL learners. The 'percentage coverage' refers to the percentage of the text that the learner is assumed to understand.

There has been continuing interest in whether there is a language knowledge threshold which marks the boundary between having and not having sufficient language knowledge for successful language use (Nation 2001). The current thinking in the field of vocabulary teaching and learning puts the threshold of meaningful input at 95% (ibid), therefore, 95% coverage was chosen as the target. Thus the percent level of each sample text vocabulary not covered by the JSH textbook should be less than 5% in order to be understood by EFL learners who studied English through these texts.

7.5 BNC text coverage

With more than 100 million words, the BNC is considered to be one of the most reliable corpus resources available, and reflects present day English usage for speech and publications in the UK. From the BNC, Chujo (2004) created a lemmatised BNC high frequency word list of 13,994 words representing 86,123,934 words in the BNC occurring 100 times or more. The words are ranked in terms of how frequently they are used, or how common they are. In teaching EFL learners to recognise spoken or written words, it is obviously important to teach them those words they are most likely to encounter. In this study, the first 1,000, 2,000, 3,000, 4,000, 5,000 most frequent words from the BNC were used as a criterion. Calculations of the percentage of words in each sample text not covered by the top 1,000, 2,000, 3,000, 4,000, 5,000 BNC words were obtained.

7.6 Japanese vocabulary ratio

The *JLPT Test Content Specifications* divides Japanese language proficiency levels into four, with Level 4 (beginner) as the first attained and Level 1 as the last (advanced). The type and size of vocabulary are specified as follows: 800 words for Level 4, 1,500 words for Level 3, 4,800 words for Level 2, and 7,800 words for Level 1. We used the Vocabulary Level Checker programme,[10] which first divides the input text sample into words using the Chasen Version 2.02 Japanese morphological analysis system (Matsumoto *et al.* 1997), then automatically compares all the words in the text with the words in the four levels of the *JLPT Test Content Specifications*, and finally shows the number of words at each level in a classification table. Kawamura (1999) demonstrated that the ratio of the sum of Levels 3 and 4 vocabulary highly correlated with text difficulty, i.e. if a text sample contained a large number of Levels 3 and 4 words, it was easier to understand; and the higher the ratio, the more easily it was understood. In this study the converse was noted; that is, in the ratio of the sum of Level 1 and Level 2 words to the total number of words in the text, calculated as the ‘Japanese vocabulary ratio’, it was noted that the higher the ratio, the more difficult the text was considered to be.

7.7 Japanese sentence length

Japanese sentence length is considered to be another measure that reflects the level of difficulty of texts. Sentence length was quantified as the average number of Japanese characters per sentence in a text, and was counted using the CL Tool programme.[11] The average sentence length was obtained by dividing the total number of characters in the text by the number of sentences. The average sentence length is also one of the numerical standards of the *JLPT Test Content Specifications*: Level 4 average sentence length falls between 20 to 25 characters, Level 3 is 25-30 characters, Level 2 is 30-45 characters, and Level 1 contains 40-65 characters.

7.8 Kanji ratio

Japanese texts consist of kanji, hiragana, katakana, and other characters such as English letters and numerals. The ratio of kanji to the total number of characters in a text was counted by using the CL Tool programme. As with vocabulary and sentence length, the ratio of kanji is also considered to be one of the measures that reflect the level of difficulty of texts (Kokusai Kouryuu Kikin 2002). For example, *JLPT Test Content Specifications* specifies that Level 4 texts contain 15-20% kanji, Level 3 20-25%, Level 2 25-35%, and Level 1 30-45%.

8. Results and discussion

The linguistic difficulty of both English and Japanese text samples measured by each of the indices is shown in Table 1. Each title is shown with the title number, the author's name, a narrative (N: white) or expository (E: grey) category, and each index score. In order to grasp the distribution of the linguistic difficulty level graphically, the titles were sorted according to the readability scores from the lowest to the highest. Also, the range for each index score was divided into three levels and colour-coded with ascending difficulty as follows: 'Level I' (white), 'Level II' (light grey), and 'Level III' (grey). As we see from Table 1, with the exception of the BNC text coverage, each index provides sufficient criteria for classifying texts into three levels according to difficulty level. Furthermore, we see that there are twenty-six titles that are broadly classified as 'Level I' in readability, average word length or JSH text coverage, indicating that at least there are some available titles that might be used at the beginner level for Japanese students of English. Of these twenty-six, sixteen are rated 'Level I' with all three indices. Since the JSH text coverage indicators are in almost all cases except one (Number 14: *The selfish giant*) well below the 95% coverage guidelines, the use of Japanese translations in parallel corpora might be helpful for Japanese junior high school students. A few of these may be useful to JFL learners, although the Japanese parallel corpus data ranked as 'Level II' for many of these titles. It is interesting to note the differences in why certain texts might be difficult, as shown by the different index scores. For example, Edgar Allan Poe's *The tell-tale heart* (Number 8) scores as 'Level I' in terms of readability and average word length, but as 'Level II' for JSH text coverage. This tells us that even though this text may be at a US fourth grade reading level, and therefore potentially accessible to Japanese junior high school students, the vocabulary currently taught in Japanese schools would not support this kind of text. Also

Table 1. Index evaluation and rating for the parallel corpora text collection

No.	Title	Author	Type of text	English indices			Japanese indices			Reference
				Readability (grade)	Word length (letters)	Not covered by JSH text (%)	Levels 1 & 2 vocabulary ratio (%)	Sentence length (characters)	Kanji character ratio (%)	Not covered by BNC top 1,000 (%)
1	Jack and the Beanstalk	J. Jacobs	N	2.4	3.7	6.4	14.1	25.2	15.8	16.1
2	The Story of the Three Little Pigs	J. Jacobs	N	2.6	3.6	7.8	7.3	24.4	14.3	17.5
3	The Happy Prince	O. Wilde	N	3.6	3.9	6.6	12.7	33.8	22.1	15.9
4	Snow White and the Seven Dwarfs	Grimm	N	4.2	4.0	8.7	11.6	27.7	17.6	15.3
5	What The Tortoise Said To Achilles	L. Carroll	N	4.3	4.1	7.0	16.2	22.7	18.3	13.2
6	The Last Leaf	O Henry	N	4.5	4.1	11.7	13.9	32.7	19.7	21.3
7	The Little Match Girl	Andersen	N	4.6	4.2	8.7	13.7	30.2	19.1	17.8
8	The Tell-Tale Heart	E.A. Poe	N	4.7	4.0	10.2	14.9	30.1	23.5	18.1
9	The Tempest	M. Lamb	N	4.7	4.1	8.8	12.4	29.0	19.7	18.4
10	The Gift of the Magi	O. Henry	N	4.8	4.1	10.0	15.4	32.6	18.5	19.7
11	A Midsummer Night's Dream (retold)	C. & M. Lamb	N	4.8	4.1	10.6	12.8	33.3	17.2	20.6
12	Hearts and Hands	O Henry	N	4.9	4.2	10.0	12.9	30.5	20.0	18.3
13	After Twenty Years	O Henry	N	5.0	4.0	9.0	15.1	28.7	18.8	15.4
14	The Selfish Giant	O. Wilde	N	5.0	4.0	3.3	10.5	32.2	22.2	12.9
15	The Adventure of the Norwood Builder	C. Doyle	N	5.1	4.1	9.0	15.9	31.4	21.4	15.6
16	A Harlem Tragedy	O Henry	N	5.1	4.1	12.3	17.6	34.2	18.8	21.1
17	Peter Pan in Kensington Gardens	J.M. Barrie	N	5.2	4.1	6.8	11.0	39.7	16.4	15.4
18	The Adventure of the Blue Carbuncle	C. Doyle	N	5.2	4.1	10.5	17.2	35.1	21.6	17.5
19	A Scandal in Bohemia	C. Doyle	N	5.3	4.1	9.6	17.8	23.1	26.4	16.7
20	A Dog of Flanders	Ouida	N	5.4	4.2	12.6	14.9	39.7	20.2	23.6

21	Dubliners	J. Joyce	N	5.5	4.1	6.1	11.1	26.6	22.3	13.8
22	The Adventure of Charles Augustus Milverton	C. Doyle	N	5.6	4.1	10.0	16.0	34.4	17.8	17.4
23	The Arrest of Arsene Lupin	M. Leblanc	N	5.7	4.3	11.1	15.9	26.6	21.6	19.2
24	To Build a Fire	J. London	N	5.7	4.2	9.6	16.9	28.8	22.9	20.1
25	The Fad of The Fisherman	G.K. Chesterton	N	5.9	4.2	8.5	16.7	26.2	23.4	17.4
26	In Midsummer Days	A. Strindberg	N	5.9	4.2	8.7	13.3	50.5	18.2	19.3
27	The Adventure of the Devil’s Foot	C. Doyle	N	6.0	4.3	9.8	17.2	35.7	24.1	17.4
28	As You Like It (retold)	M. Lamb	N	6.0	4.2	10.5	14.4	36.4	21.6	18.3
29	Romeo and Juliet (retold)	C. & M. Lamb	N	6.3	4.3	12.7	15.4	44.5	21.0	20.6
30	The Shadow and the Flash	J. London	N	6.5	4.4	12.3	17.6	34.3	21.3	21.7
31	A Tale about a Queer Client	C. Dickens	N	6.9	4.3	11.3	18.7	37.3	26.4	21.6
32	A Horseman in the Sky	A. Bierce	N	7.2	4.4	12.3	20.3	33.9	26.3	23.4
33	An Imperfect Conflagration	A. Bierce	N	7.3	4.3	14.0	16.4	40.8	25.6	19.8
34	The Manager FAQ	P. Seebach	E	7.4	4.5	13.4	17.8	37.0	21.0	15.3
35	The Assignation	E.A. Poe	N	7.7	4.6	14.7	16.8	51.2	25.6	25.2
36	Clinton’s Inaugural Address	B. Clinton	E	7.7	4.4	10.4	24.6	41.2	25.2	18.6
37	The Black Cat	E.A. Poe	N	7.9	4.4	18.1	15.9	49.2	22.9	23.0
38	The Man and the Snake	A. Bierce	N	7.9	4.5	14.4	20.7	36.9	28.6	23.9
39	Waiting for the Knock	R.M. Stallman	E	8.0	4.5	11.2	21.6	42.6	23.7	16.5
40	JFK’s Inaugural Address. 1/20/1961	R.M. Stallman	E	8.0	4.3	11.1	24.8	47.5	27.8	21.3

		White:	Level I
N:	Narrative	Light Grey:	Level II
E:	Expository	Grey	Level III

Table 1. (Cont.)

41	Someone to watch over me	J.F. Kennedy	E	8.0	4.5	9.0	20.0	52.2	23.6	9.5
42	A Biographical Sketch of an Infant	C. Darwin	E	8.1	4.3	8.8	16.1	43.6	26.8	16.8
43	The Manager FAQ	Peter Seebach	E	8.2	4.4	7.6	19.2	43.3	25.7	10.1
44	RMS Lecture at KTH (Sweden). 1986	R.M. Stallman	E	8.4	4.2	5.9	15.8	47.1	16.8	10.5
45	Boycott Amazon!	R.M. Stallman	E	8.5	4.6	11.8	22.6	49.4	23.2	15.9
46	Interview with Karl Marx	R. Landor	E	8.7	4.4	11.6	25.1	43.8	25.5	16.4
47	A Sense of History: Some Components	G.W. Schlabach	E	8.8	4.8	10.4	19.8	34.9	27.8	16.3
48	Linux and GNU-GNU Project	R.M. Stallman	E	8.8	4.6	10.5	20.4	55.1	15.0	10.7
49	A Modest Proposal for preventing the children ...	J. Swift	E	9.1	4.5	15.0	21.0	44.7	32.1	21.2
50	Twenty rules for writing detective stories	S.S. Van Dine	E	9.5	4.6	16.8	26.3	38.5	33.4	27.5
51	The Pragmatist of Free Software	Hiroo Yamagata	E	9.6	4.3	8.3	17.4	37.9	17.0	13.0
52	Why “Free Software’’ is better than “Open Source’’	R.M. Stallman	E	9.6	4.7	9.6	27.0	55.0	20.8	15.6
53	Some Confusing or Loaded Words …	FSF	E	10.0	4.9	16.6	27.6	55.3	27.2	20.0
54	The Abolition of Work	B. Black	E	10.1	4.9	16.6	25.8	44.3	31.4	21.5
55	Free Software is More Reliable!	FSF	E	10.1	4.9	19.2	24.3	62.6	16.8	24.8
56	The Fall of the House of Usher	E.A. Poe	N	10.2	4.6	18.4	19.9	63.0	23.3	26.4
57	Freedom-Or Copyright?	R.M. Stallman	E	10.2	5.0	16.0	24.4	42.5	27.4	21.2
58	The Darwinian Hypothesis	T.H. Huxley	E	10.4	4.7	14.7	23.6	68.7	30.5	21.4
59	The Beginning of Ownership	T. Veblen	E	11.2	4.8	18.0	28.2	63.4	36.5	21.7
60	Why you should use the GNU FDL	R.M. Stallman	E	11.5	5.1	18.8	29.6	56.7	23.1	22.5

61	The Declaration of Independence	USA.	E	11.6	4.9	17.6	27.5	61.7	34.2	23.1
62	Discourse on the Method of Rightly …	R. Descartes	E	12.3	4.5	11.3	20.4	69.5	23.3	17.7
63	The Yomiuri Shimbun/The Daily Yomiuri #1		E	12.6	5.0	16.5	32.3	51.6	46.2	21.8
64	The Yomiuri Shimbun/The Daily Yomiuri #2		E	13.0	5.1	15.4	31.2	53.0	45.2	21.2
65	The Instinct of Workmanship …	T. Veblen	E	13.5	5.0	15.3	31.9	55.9	44.5	19.7
66	The Yomiuri Shimbun/The Daily Yomiuri #3		E	13.5	4.8	17.0	26.8	58.1	35.8	21.3
67	The Yomiuri Shimbun/The Daily Yomiuri #4		E	13.6	5.1	17.3	32.3	53.2	46.1	21.2
68	The Yomiuri Shimbun/The Daily Yomiuri #5		E	13.7	5.1	16.5	32.0	56.7	45.3	20.6
69	The Yomiuri Shimbun/The Daily Yomiuri #6		E	13.8	5.1	18.1	32.8	49.5	45.4	22.0
70	The Yomiuri Shimbun/The Daily Yomiuri #7		E	13.8	5.1	16.8	31.1	52.8	44.3	20.5
71	The Yomiuri Shimbun/The Daily Yomiuri #8		E	14.0	5.1	17.3	32.0	46.9	45.3	21.4
72	The Yomiuri Shimbun/The Daily Yomiuri #9		E	14.0	5.1	15.5	32.5	53.4	45.1	20.6
73	The Yomiuri Shimbun/The Daily Yomiuri #10		E	14.1	5.1	16.1	30.5	53.2	46.1	20.9
74	The Yomiuri Shimbun/The Daily Yomiuri #11		E	14.3	5.3	16.1	36.1	39.8	44.3	22.7
75	The Yomiuri Shimbun/The Daily Yomiuri #12		E	15.1	5.3	16.0	36.9	47.2	42.2	21.1

		White:	Level I
N:	Narrative	Light grey:	Level II
E:	Expository	Grey	Level III

interesting is that although former US president Bill Clinton's inaugural address (Number 36) is rated as 'Level II' in readability, average word length and JSH text coverage, the Japanese parallel text ranks as 'Level III' for both vocabulary and kanji. As we expected, each index for the twelve newspaper text samples shows scores distributed within a narrow range, indicating they are fairly uniform in their difficulty level, and that almost all indices rank the newspaper texts in one classification ('Level III'). From this we can predict that these newspaper texts might provide a stable corpus for advanced level learners.

Surprisingly, the calculations for text coverage for the top 1,000 to 5,000 BNC words produced inconsistent scores. Only the top 1,000 BNC scores are displayed in Table 1 as a reference to illustrate what scores were produced, and how they compared to the other indices.

The criteria for dividing these data into three levels by each index are shown in Table 2. A detailed discussion on each index result follows.

Table 2. Overview of level definitions

Level	English			Japanese			Reference
	Readability	**Word length**	**Percentage not covered by JSH vocabulary**	**Levels 1 & 2 vocabulary ratio**	**Sentence length**	**Kanji character ratio**	**Percentage not covered by BNC top 1,000**
Level I	Lower than US grade 5.9 (JPN Junior HS Level)	Shorter than 4.2 letters	Less than 9.9%	Less than 15.9%	Fewer than 30.4 characters	Fewer than 19.9%	Less than 16.9%
Level II	Between US grades 6.0 & 8.9 (JPN Senior HS Level)	Between 4.3 & 4.5 letters	Between 10 & 14.9%	Between 16 & 21.9%	Between 30.5 & 44.9 characters	Between 20 & 24.9%	Between 17.0 & 20.9%
Level III	Higher than US grade 9.0 (JPN College Level)	Longer than 4.6 letters	More than 15%	More than 22%	Longer than 45 characters	More than 25%	More than 21%

8.1 English readability findings

The English readability score is shown in the 5th column of Table 1. This data was classified into three rankings, based on the US reading grade and the corresponding Japanese English school textbook readability score: 'Level I' (lower than the US 5.9 grade level, i.e. at the Japanese junior high school English textbook level), 'Level II' (between the US 6.0 and 8.9 grade levels, i.e. at the

Japanese senior high school textbook level), and ‘Level III’ (higher than the US 9.0 grade level, i.e. at the Japanese college level or beyond). The non-shaded (white) scores indicate that the texts are at the appropriate level for Japanese junior high school students, and the light-grey scores are titles appropriate for senior high school students. It was noted that 15 titles and all of the newspaper texts were higher than the 9.0 grade level, or higher than the Japanese senior high school graduate level. Out of these 15 titles, 14 were expository texts. Reading comprehension research tells us texts can be defined as either narrative (a fiction or non-fiction story) or expository (non-narrative, an explanation or source of information). Generally, narrative texts are easier to comprehend than expository texts, since

> … [n]arratives possess a well-documented, familiar structure [and] … from a content perspective, narratives typically deal with information about social or interpersonal relationships and everyday problem solving, content about which both adults and children tend to know quite a bit. (Cote 1998: 6)

It is not surprising then that we see in Table 1 that many of the readability scores for expository material (newspaper articles, academic papers, commentaries, blogs and political speech texts) are higher than the 9th grade level, and beyond the reach of Japanese senior high school students:

> Compared to narratives, expository text structures are more variable. … [A] common purpose of expository texts is informational. Informational texts frequently present concepts and relations that readers do not already know. They require that readers understand a greater range of logical relations among pieces of information. (ibid)

Thus, understanding expository texts generally requires more knowledge from readers, and consequently, is generally considered difficult by educators.

8.2 Average word length findings

The average word length of each text sample is shown in the 6th column of Table 1. The average number of letters was calculated, yielding a range from 3.6 to 5.3. Studies such as Chujo and Takefuta (1989), and Takefuta, Hasegawa and Chujo (1994) showed that the longer the word, the higher the level of difficulty. Word length was also used as one of the variables to classify texts in Biber (1988). In Table 2, based on Chujo and Takefuta (1989), words shorter than 4.2 letters were defined as ‘Level I’, and those between 4.3 and 4.5 letters were defined as ‘Level II’. Those longer than 4.6 letters were defined as ‘Level III’. As we might expect, most narratives were evaluated as ‘Level I’ and most expository texts were defined as ‘Level III’ in terms of average word length.

8.3 JSH text coverage findings

The percentage of vocabulary not covered by the JSH textbooks is shown in the 7th column. This JSH textbook vocabulary represents the vocabulary a learner usually acquires before entering a university. Researchers such as Laufer (1992) and Nation (2001) pointed out that learners would need 95% text coverage to understand the meaning of texts. It turns out that only one text, *The selfish giant* (Number 14), fulfilled this criterion with a score of 3.3%. From this we can easily imagine that those Japanese learners who studied English solely through school English textbooks would have difficulty reading the English concordance lines of any of the titles in the Collection except those from this one book. This result shows that not only is the validity of Japanese textbooks (and their vocabulary selection) called into question, but that, in spite of gaps in vocabulary learning in Japanese schools, the use of parallel concordancing lines showing Japanese translations would be not only helpful to learners, but perhaps essential. Teachers using the 'Level I' titles shown in white under the JSH column, for example, might have greater success if students also used the Japanese concordancing lines since much, but not 95%, of the vocabulary in these titles is covered in the Japanese textbooks. It is also important to note that while the 95% coverage is the ideal, the 'Level I', 'Level II' and 'Level III' guidelines in the study were set respectively as 9.9%, between 10.0% and 14.9%, and above 15%. These divisions were created in five point intervals within the distribution of the range of JSH text coverage obtained from the titles. From Table 1, we see that the ratio of unknown words is greater in expository texts than in narrative texts, which follows similar observations in the readability and word length indices.

8.4 BNC text coverage findings

The BNC top 1,000 text coverage is shown in the far most right column. Since the BNC represents present day general vocabulary usage, we expected that the BNC high frequency word lists would function as an appropriate tool for measuring vocabulary levels of the various texts investigated in this study. Contrary to our expectation, the BNC lists were found not to correlate significantly with the other indices. As you can see from Table 1, there is an inconsistent variation on text coverage both between and within the 'Level I', 'Level II' and 'Level III' text samples, indicating that BNC coverage is not a similar predictor of level compared to the other six indices. While we believe that BNC text coverage is a stable index, it may be that the BNC's rating differences are a factor of genre or publication dates in a way that the other indices are not. For example, the BNC would rate blogs as 'Level I' and novels as 'Level II' or 'Level III'. If we accept that the BNC is a corpus of present day usage, this makes sense since blogs, a very modern invention, are generally written with informal language (sometimes in the style of personal journals), and novels usually use more formal or descriptive vocabulary.

So while at first glance the BNC did not seem to provide much useful data for this study in terms of being a similar index to the others used, the data obtained does raise interesting questions with regard to text selection, i.e. the type of text (spoken, informative or imaginative) and its degree of modernity as a factor in how accessible it will be to learners. These questions warrant further study, but because BNC text coverage was not a clear and consistent indicator of level regarding these specific texts, for the purposes of this study it was excluded from additional between-index calculations. Also note that, as with JSH classification, there was some degree of arbitration in defining the 'Level I', 'Level II' and 'Level III' categories. For the percentage of top 1,000 BNC text coverage, these were chosen from the distribution obtained (9.5% to 27.5%): those less than 16.9% were defined as 'Level I,' those between 17.0 % and 20.9% as 'Level II', and those more than 21% as 'Level III'.

8.5 Japanese vocabulary ratio findings

The ratio of Levels 1 and 2 vocabulary specified in the *JLPT Test Content Specifications* is shown in the 8th column of Table 1. There is no published standard to classify the 'Japanese vocabulary ratio', so these were set within the distribution obtained (7.3%-36.9%) as follows: those less than 15.9% were defined as 'Level I', those between 10.0% and 14.9% as 'Level II', and those more than 22.0% as 'Level III'.

The ratio of this advanced level vocabulary in narrative texts is roughly about 15%, and is approximately 23% for expository texts except for newspaper samples. Not unexpectedly, newspaper articles are at the 32% level. This confirms the belief that expository texts use more difficult vocabulary than narrative texts. This is a valuable finding for JFL learners using corpora.

8.6 Japanese sentence length findings

Japanese sentence length is shown in the 9th column in Table 1. The texts can be classified into three groups based on the *JLPT Test Content Specifications*. The titles which contain an average sentence length of fewer than 30.4 characters are defined as 'Level I', and correspond to JLPT Levels 3 and 4. Titles having a sentence length between 30.5 and 44.9 characters are defined as 'Level II', and correspond to JLPT Level 2. Those titles with an average sentence length longer than 45.4 characters are defined as 'Level III'; these correspond to JLPT Level 1. Measuring Japanese average sentence length also clearly shows that expository texts use longer sentences. Many of the expository text sentences are longer than 50 characters per sentence.

8.7 Kanji ratio findings

The kanji character ratio is shown in the 10th column in Table 1. Titles can be broadly classified into three groups according to the kanji ratio criteria shown in Table 2. Again based on the *JLPT Test Content Specifications*, titles having a kanji ratio fewer than 19.9% (JLPT Level 4) are rated as 'Level I', between 20.0 and 24.9% (JLPT Level 3) are 'Level II', and more than 25.0% (JLPT Levels 1 and 2) are 'Level III'.

Kanji creates a huge learning burden for JFL learners or even Japanese whose kanji proficiency level is low. Unfortunately for these learners, this study showed that half of the Japanese texts investigated contained JLPT Level 1 or 2 kanji. Not surprisingly, newspapers in particular use a substantial amount of kanji: on average, 45% of the text.

8.8 Correlation among indices

All of the above observations of the indices provided valuable insight regarding the linguistic difficulty levels. Each, with the exception of the BNC,[12] was a sufficient criterion for classifying the particular texts investigated in this study. We acknowledge that there was some degree of arbitration in defining difficulty levels for the indices, so in order to understand the extent of the similarity between indices regarding the Level I, II and III classifications, we next calculated the correlation between indices.

Table 3 shows the results for English texts and Table 4 for Japanese texts, with high correlations (greater than 0.75) shaded to enhance clarity. Averaged correlation coefficients between indices were shown in each bottom row.

Table 3. Correlation between indices for English texts

	Readability	**Word length**	**JSH text coverage**
Readability	—	0.95	0.76
Word length	0.95	—	0.82
JSH text coverage	0.76	0.82	—
Average	0.85	0.88	0.79

▨ > 0.75

Table 4. Correlation between indices for Japanese texts

	Levels 1 & 2 vocabulary ratio	**Sentence length**	**Kanji character ratio**
Levels 1 & 2 vocabulary ratio	—	0.64	0.86
Sentence length	0.64	—	0.47
Kanji character ratio	0.86	0.47	—
Average	0.75	0.56	0.67

▨ > 0.75

All three indices for the English texts showed a high correlation to each other, and this correlation provides some support for the degree of arbitration used in the study to define levels of difficulty. We see the strongest correlation between readability and word length (0.95). Using the averaged correlation coefficient as a means of comparison, the word length index showed the highest correlation with other two English indices (0.88); thus we can say that this index is regarded to be the most effective index among three indices observed in this study in classifying English text difficulty. However, all three indices for English texts showed a high correlation to each other, indicating that all three indices are effective. Of these, the word length index will probably be the most easy-to-use criterion for teachers since it can be obtained in one step with readily available software such as Writers' Workbench. Teachers in Japan evaluating texts for junior and senior high school students may also want to consider JSH percentages since, while the indices can and do reasonably target level, the percentage of coverage (or lack thereof) is a useful guideline unless students have the use of Japanese parallel text data to compensate for the lack of known vocabulary.

We see the strongest correlation between Japanese Levels 1 and 2 vocabulary ratios and kanji character ratios (0.86) for Japanese texts. This would follow conventional wisdom that higher level vocabulary would be expressed in kanji. Looking at the averaged correlation coefficient as a means of comparison, Japanese Levels 1 and 2 vocabulary ratios showed the highest correlation with the other two Japanese indices (0.75); thus this index is regarded to be the most useful index among three indices observed in this study in classifying Japanese text difficulty. This ratio can be calculated easily by using Vocabulary Checker software available on the web (see Kawamura 1999).

Next, in order to explore the possibility that the English texts' difficulty level correlates with its Japanese translation texts' difficulty level, for, in fact, they are saying the same thing, we calculated the correlation between pairs of English and Japanese indices in Table 5. The three Japanese indices are on the top row and the three English indices are in the first column. In the far right column, the averaged correlation between an English index and each of the three Japanese indices are shown. In the bottom row, the averaged correlation between a Japanese index and each of the English indices are shown.

Table 5. Comparing parallel English and Japanese text difficulty by indices

	Levels 1 & 2 vocabulary ratio	**Sentence length**	**Kanji character ratio**	**Average**
Readability	0,93	0,77	0,85	0,85
JSH text coverage	0,76	0,66	0,66	0,69
Word length	0,93	0,71	0,82	0,82
Average	0,87	0,71	0,78	

▨ > 0.75

Overall, there is a clear correlation between the English and Japanese indices. It was interesting that English readability had the highest correlation average with all three indices for Japanese texts, followed by word length. Japanese Levels 1 and 2 vocabulary had the highest correlation average with all three indices for English texts, followed by kanji ratio. We might infer from this that it might be possible to gauge the difficulty level of Japanese translation texts from the English counterpart's readability, and vice versa. For example, when English texts are expository, it is not surprising that the Japanese translations would use advanced Japanese vocabulary, longer sentences and a larger number of kanji.

9. Conclusion

In this study, the linguistic difficulty of English and Japanese text samples taken from two parallel corpora were measured with seven indices. Six were shown to be applicable, and reliability was demonstrated by correlations between indices. A text collection was created and the titles have been listed with each of the seven indices (readability, word length, JSH textbook vocabulary coverage, BNC text coverage, kanji ratio, Japanese vocabulary level and Japanese sentence length) to define specifically in what way they are difficult. It was found that the best and easiest way to evaluate the level of difficulty for English texts is by using average word length, although readability scores and JSH textbook coverage can also be useful. For measuring the level of difficulty in Japanese texts, vocabulary was the most effective. Also, the level of difficulty for one language was generally a reliable predictor for the difficulty of the parallel language texts, i.e. a 'Level I' English text generally had a 'Level I' Japanese translation, and a 'Level III' English text corresponded to a 'Level III' Japanese translation, and vice versa. Not surprisingly, it was found that many expository texts and all the newspaper articles were difficult in both languages. The BNC text coverage data did not provide comparable results but did raise interesting questions about the type (spoken, informative or imaginative) and modernity of texts chosen for learners, especially those at the beginner level.

There is an unfortunate shortage of copyright available e-text material at the beginner level, and the readability comparisons between American and Japanese grades discussed in this study might point to American graded readers as a potential source of corpus data. In addition to the necessity for finding and including more beginner level corpus data, it is also clear that the vocabulary taught in Japanese junior and senior high schools may need review. In this study, only one 'Level I' title contains vocabulary understood by the average high school graduate at a 95% coverage level. Perhaps it is time for the Japanese Ministry of Education, Science and Culture to create scientific and modern vocabulary guidelines based on recent work in corpus linguistics.

A logical extension for future study would be to add easier level reading texts to the text collection, to compare each index with the subjective difficulty level standards obtained from educators, and to quantify and measure how much the difficulty level of English texts will be reduced by the use of Japanese translations. In the meantime, the use of Japanese parallel concordancing lines might be a reasonable tool in understanding English texts, and a case study on this topic is the subject of our next research project.

Notes

[1] English texts were mainly collected from Project Gutenberg (at «http://promo.net/pg/») and the GNU Project (at «http://www.gnu.org/»), with Japanese translations from Project Sugita Genpaku (at «http://www.genpaku.org/») and from other resources. The project (Utiyama 2003) is on-going and the corpus is available at «www2.nict.go.jp/jt/a132/members/mutiyama/align/index.html».

[2] This corpus was created by Utiyama and Isahara (2003) and is available at «www2.nict.go.jp/jt/a132/members/mutiyama/jea/».

[3] For a discussion on how the English texts were prepared for readability and text coverage calculations, see Chujo *et al.* (2004).

[4] The following textbook series were used: Asano *et al.* (1999) and Suenaga *et al.* (2001).

[5] Vocabulary Checker available at «http://language.tiu.ac.jp/tools.html» (see Kawamura 1999).

[6] Micro Power & Light Co. 2003: Readability Calculations. «http://www.micropowerandlight.com».

[7] These formulas use one or more of the following criteria to calculate the score: number of words, number of syllables, number of sentences,

average number of syllables, and number of words more than three syllables.

[8] EMO Solutions 2004: Writer's Workbench Version 8.15 «http://www.emo.com/wwb/».

[9] We used our own programme for calculating text coverage, but a similar programme is available at Paul Nation's web site «http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx».

[10] It is worth noting that this software is only capable of handling 19KB of data (about 200 sentences), so two random samples of 200 sentences from each title were extracted and averaged together to calculate the score.

[11] CL TOOL Version 1.2 «http://sano.tufs.ac.jp/cltool/» (see Sano 2003).

[12] The correlation of the BNC top 1,000 percentages with readability, word length and JSH text coverage was .37, .45, and .77 respectively. Its correlation with Japanese vocabulary, sentence length and kanji ratio was .41, .32, .43 respectively.

References

Asano, H., Y. Shimomura, T. Makino, M. Ikeda, A. Ikeya and S. Ishizuya (1999), *New Horizon English Course 1, 2, and 3*. Tokyo: Tokyo Shoseki.

Aston, G. (2001), *Learning with corpora*. Houston: Athelstan.

Biber, D. (1988), *Variations across speech and writing*. Cambridge: Cambridge University Press.

Chujo, K. (2004), 'Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list', in: J. Nakamura, N. Inoue and T. Tomoji (eds.), *English corpora under Japanese eyes*. Amsterdam: Rodopi, 231-249.

Chujo, K. and Y. Takefuta (1989), 'Joseimuke Eigo zasshi no goi (vocabulary of women's magazines)', *Current English Studies*, 28: 73-84.

Chujo, K., A. Shirai, M. Utiyama, C. Nishigaki and S. Hasegawa (2004), 'Nichiei parallel corpus wo kouseisuru text no nan'ido ni kansuru kenkyuu (a study on classifying texts in English-Japanese parallel corpora according to linguistic difficulty)', *Journal of the College of Industrial Technology Nihon University*, 37: 57-68.

Cote, N., S.R. Goldman and E.U. Saul (1998), 'Students making sense of informational text: relations between processing and representation', *Discourse Processes*, 25: 1-53.

Flesch, R. (1974), *The art of readable writing*. New York: Harper and Row.

Fry, E. (1968), 'A readability formula that saves time', *Journal of Reading*, 11/7: 265-271.

Kawamura, Y. (1999), 'Analysis of Japanese textbooks using the "Vocabulary Level Checker"', in: K. Nakajima *et al.* (eds.), *Proceedings of the Second*

International Conference on Computer Assisted System for Teaching & Learning Japanese. Toronto: University of Toronto, 132-137 (retrieved from «http://language.tiu.ac.jp/about.html»).

Kokusai Kouryuu Kikin (2002), *Nihongo Nouryoku Shiken Shutsudai Kijun* (Japanese Language Proficiency Test: Test Content Specifications) (Revised Edition), Tokyo: Bonjinsha.

Laufer, B. (1992), 'How much lexis is necessary for reading comprehension?', in: L. Arnaud and H. Bejoint (eds.), *Vocabulary and applied linguistics*. London: Macmillan, 126-132.

Matsumoto, Y., K. Kitauchi, T. Yamashita, O. Imaichi and T. Imamura (1997), 'Nihongo keitaiso system Chasen manual version 1.0'. *NAIST Technical Report* 97007. (Version 2.3.3 is available at «http://chasen.aist-nara.ac.jp/chasen/manual.html.ja») .

McLaughlin, G. (1969), 'SMOG grading: a new readability formula', *Journal of Reading*, 12/8: 639-646.

Nation, P. (2001), *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Sano, H. (2003), *Windows PC niyoru Nihongo kenkyuuhou* (methods of investigating Japanese language by use of Windows PC). Tokyo: Kyoritsu Shuppan Co. Ltd.

Suenaga, K., Y. Yamada, K. Fukai, S. Nakamura, K. Ishizuka and K. Ichinose (2001), *Unicorn English Course I, II, and Reading*. Tokyo: Bun'eido.

Takefuta, Y., S. Hasegawa and K. Chujo (1994), 'Goi list "Gendaieigo no Keyword" no nin'chi level niyoru kubun no datousei (validity of cognitive level grading for Keyword System 5000)', *Working Papers in Language and Speech Science*, 4: 53-63.

Thomas, J. (2002), 'Concordancing for all', paper presented at the *Fifth Teaching and Language Corpora Conference*, Bertinoro, Italy, 27-31 July 2002.

Tono, Y. (2003), 'Corpus wo Eigo kyoiku ni ikasu (what corpora can do for language teaching)', *English Corpus Studies*, 10: 249-264.

Utiyama, M. (2003), 'Japanese-English bilingual corpora and their applications', demonstration presented at the *Asialex Conference*, Tokyo, Japan, 27-29 August 2003.

Utiyama, M. and H. Isahara (2003), 'Nichiei shimbun no kiji oyobi bun wo taiouzukeru tameno koushinraisei shakudo (reliable measures for aligning Japanese-English news articles and sentences)', *Journal of Natural Language Processing*, 10/4: 201-220.