Text Readability and Word Distribution in Japanese

Satoshi Sato

Nagoya University
Chikusa-ku, Nagoya, 464-8603, JAPAN
ssato@nuee.nagoya-u.ac.jp

Abstract

This paper reports the relation between text readability and word distribution in the Japanese language. There was no similar study in the past due to three major obstacles: (1) unclear definition of Japanese "word", (2) no balanced corpus, and (3) no readability measure. Compilation of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and development of a readability predictor remove these three obstacles and enable this study. First, we have counted the frequency of each word in each text in the corpus. Then we have calculated the frequency rank of words both in the whole corpus and in each of three readability bands. Three major findings are: (1) the proportion of high-frequent words to tokens in Japanese is lower than that in English; (2) the type-coverage curve of words in the difficult-band draws an unexpected shape; (3) the size of the intersection between high-frequent words in the easy-band and these in the difficult-band is unexpectedly small.

Keywords: Japanese vocabulary, type-coverage curve, readability

1. Introduction

Vocabulary research is the foundation of language technology. In Japan, the National Institute of Japanese Language and Linguistics (NINJAL) conducted a series of vocabulary studies from 1950s and the results were used in applications such as compilation of Japanese dictionaries and vocabulary lists for education.

Frequency is one of the most important characteristics of vocabulary (Schmitt, 2010). In order to obtain reliable frequency data, a large and well-designed balanced corpus is required. Such a corpus for Japanese was not compiled before 2000.

The Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2010) released in 2012 is the first *balanced* corpus of Japanese with 100 million words. This corpus is suitable to vocabulary research because it provides the word-segmented texts in addition to the row texts.

In an agglutinative language such as Japanese, the definition of *word* is not clear due to no obvious boundary such as a white space in English texts. This causes a serious problem when we count frequencies of words. BCCWJ solves this problem by employing two different levels of *word* and providing the detailed definition of them (Ogura et al., 2011a; Ogura et al., 2011b).

This paper reports the result of a study that examines the relation between text readability and word distribution in the Japanese language. There was no similar study in the past due to three obstacles: (1) unclear definition of Japanese *word*, (2) no balanced corpus, and (3) no readability measure. First two obstacles have been removed by compilation of the above mentioned BCCWJ. The last obstacle has been removed by development of a Japanese readability predictor.

The rest of this paper is organized as follows. Section 2 describes the corpus used in this study and Section 3 describes the Japanese readability predictor. Three sections from Section 4 describe three major findings in this study.

2. The Corpus

This study uses a part of the Balanced Corpus of Contemporary Japanese (BCCWJ) (Maekawa et al., 2010), which was compiled by the National Institute for Japanese Language and Linguistics (NINJAL). BCCWJ is the first *balanced* corpus of Japanese with 100 million words.

BCCWJ has two types of samples, which are extracted for a randomly-sampled maker in a randomly-sampled document.

a *fixed-length sample* consists of 1,000 characters just after the maker in the document.

a *variable-length sample* is a discourse unit (such as chapter and section) less than 10,000 characters that contains the maker.

This study uses the fixed-length samples because they are designed for statistical analysis.

As mentioned before, the definition of *word* in Japanese is not clear. In fact, a dozen different definitions of *word* have been proposed in Japanese linguistics. In order to fix this vagueness, BCCWJ employs two different levels of *word*:

a *short-unit word* (SUW) corresponds to a simple and short word that has no internal structure.

a *long-unit word* (LUW) is a sentential component, which consists of one or more SUWs.

For every samples, BCCWJ provides a result of the word-segmentation analysis, where every SUW and LUW is identified with a part of speech.

Table 1 shows an analysis example of the Japanese phrase of "国立国語研究所においては (at National Institute of Japanese Language and Linguistics)". This phrase consists of eight SUWs, which correspond to three LUWs. In further sentence analysis such as parsing, a LUW can be viewed as a single component.

NINJAL does not provide any readability score of each sample in BCCWJ. We have assigned a readability score

Japanese	国立	国語	研究	所	に	おい	て	は
GLOSS	national	language	research	institute	CASE	regarding	CASE	TOPIC
SUW	noun	noun	noun	suffix	particle	verb	particle	particle
LUW	noun (compound)				particle (compound)			particle

Table 1: Analysis Example (from (Maekawa et al., 2010))

			total				average per sample			
			SUW		LUW		SUW		LUW	
corpus	readability	sample	token	type	token	type	token	type	token	type
A	1	847	493,892	21,351	429,803	35,098	583.1	224.2	507.4	217.8
	2	1,401	860,712	27,788	771,929	43,442	614.4	239.2	551.0	235.2
	3	3,169	2,009,257	45,704	1,763,974	89,024	634.0	248.0	556.6	240.9
	4	3,342	2,122,563	53,632	1,797,204	119,733	635.1	252.2	537.8	239.7
	5	4,341	2,749,772	64,293	2,233,361	171,613	633.4	250.3	514.5	233.3
	6	2,796	1,798,372	48,648	1,387,425	132,458	643.2	248.2	496.2	228.5
	7	2,190	1,382,373	33,164	1,041,056	103,493	631.2	229.7	475.4	215.5
	8	1,502	896,434	20,202	653,664	74,821	596.8	204.0	435.2	193.7
	9	956	649,531	12,378	456,108	43,405	679.4	200.5	477.1	194.7
	(total)	20,544	12,962,906	107,243	10,534,524	515,203	631.0	240.2	512.8	227.8
B	1	800	466,651	20,741	406,334	33,813	583.3	224.3	507.9	217.9
	2	800	490,547	21,051	440,263	30,104	613.2	238.9	550.3	235.2
	3	800	505,958	24,560	444,145	36,457	632.4	248.7	555.2	241.1
	4	800	508,479	27,669	430,718	44,096	635.6	252.1	538.4	240.0
	5	800	507,347	29,647	411,689	50,573	634.2	250.6	514.6	233.3
	6	800	515,258	26,436	398,456	51,337	644.1	247.6	498.1	228.4
	7	800	503,034	21,021	379,295	48,430	628.8	229.8	474.1	215.5
	8	800	478,803	15,376	349,742	45,840	598.5	204.3	437.2	194.6
	9	800	543,033	11,516	381,724	38,234	678.8	200.4	477.2	194.7
	(total)	7,200	4,519,110	69,135	3,642,366	235,836	627.7	232.9	505.9	222.3
	Easy (1–3)	2,400	1,463,156	38,639	1,290,742	71,617	609.6	237.3	537.8	231.4
	Moderate (4–5)	2,400	1,531,084	48,592	1,240,863	110,636	638.0	250.1	517.0	233.9
	Difficult (6–9)	2,400	1,524,870	28,599	1,110,761	108,335	635.4	211.5	462.8	201.6

Table 2: Corpus-A and Corpus-B

to it by Obi2/B9 system (Sato, 2011). The score is an integer between one and nine, which corresponds to a *relative* readability level. The detail will be described in the next section.

Table 2 shows two sub-corpora that we have actually used. Corpus-A consists of almost all fixed-length *book* samples, excluding 124 exceptionals. The number of samples with each readability score varies in Corpus-A. In contrast, Corpus-B, which is a part of Corpus-A, consists of 7,200 samples, 800 samples for each of nine readability scores.

3. Readability Score

Obi2 is a language-model-based Japanese readability predictor[1], a new version of the Obi1 system reported in (Sato et al., 2008). The difference of Obi2 from Obi1 is the language model: Obi2 uses the character-bigram model instead of the character-unigram model.

A training corpus of a language model for readability prediction is called a *criterion*. A criterion consists of several sub-corpora, each of which corresponds to a readability level or score. Obi1 uses the textbook corpus (Sato et al., 2008), which consists of 1,478 sample texts extracted from 127 textbooks in thirteen school grades in Japan: elementary school (1–6), junior high school (7–9), high school (10–12), and college (13). Obi2/T13, the predictor that uses this criterion, produces an integer between 1 and 13, which corresponds to a Japanese school grade. T13, which we call a *scale* of readability measurement, is an *absolute* scale.

Obi2 has another scale, B9, which is a *relative* scale. The criterion of the B9 scale is the corpus that consists of variable-length samples (18,000 *book* samples) in BCCWJ. Because NINJAL does not provide any readability score of each sample in BCCWJ, we have assigned it by the following method (Sato, 2011). First, we sorted all samples in readability order, i.e., from easiest to the most difficult, by using a similar method proposed by (Tanaka-Ishii et al., 2010). Then, we assigned a score to each sample according to the stanine, shown in Figure 1. The distribution of the readability scores follows the normal distribution with a mean of five and a standard deviation of two. Obi2/B9 is the predictor that uses this criterion, which produces an integer between 1 to 9; the score indicates the percentile of the examined text in the readability distribution of Japanese books.

Finally, we assigned a readability score to every fixed-length samples and re-assigned it to every variable-length

[1] http://kotoba.nuee.nagoya-u.ac.jp/sc/obi2/

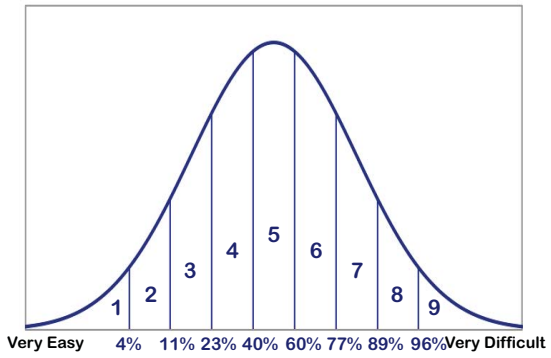

Figure 1: Stanine

r	SUW	LUW
75%	987	1,649
80%	1,736	3,509
85%	3,108	8,396
87%	4,024	12,473
90%	6,165	24,323
95%	14,860	98,314

Table 3: The number of types that requires to cover r% tokens

samples in BCCWJ by using Obi2/B9. The reliability of the assigned scores has been confirmed by experiments with human subjects (Sato and Kashino, 2012).

4. Type-Coverage Curves

Figure 2 shows the type-coverage curves of SUWs and LUWs in Corpus-A. In this graph, the x-axis represents the number of word types in the frequent order. The y-axis represents the coverage, which is the proportion of the accumulative number of tokens of high-frequent words to the total number of tokens in the corpus. From this graph, we can see a big difference between two curves. Table 3 shows required numbers of types for some typical percent coverages. Note that the required number of types is larger than that of English; it has been reported that 2,000 high-frequent English words cover 87% of tokens (Nation, 1990). In case of Japanese, 4,024 SUWs are required to cover 87% of tokens.

5. Type-Coverage Curves of Three Readability Bands

An interesting result has been observed when we have drawn the type-coverage curve for each of three readability bands of Corpus-B: **E**asy-band (the readability score is between 1 and 3), **M**oderate-band (begween 4 and 6), and **D**ifficult-band (begween 7 and 9). Our expectation was that the coverage curve of an easier band would move up more rapidly than that of more difficult band. In case of LUWs, this expectation is confirmed, shown in Figure 3. In case of SUWs, however, it is not, shown in Figure 4.

Figure 5, which is an enlargement of Figure 4, shows the coverage curves of the top 2,000 frequent SUWs. In the top 200 frequent words, the coverage of the D-band is the lowest among three bands. However, the coverage of the

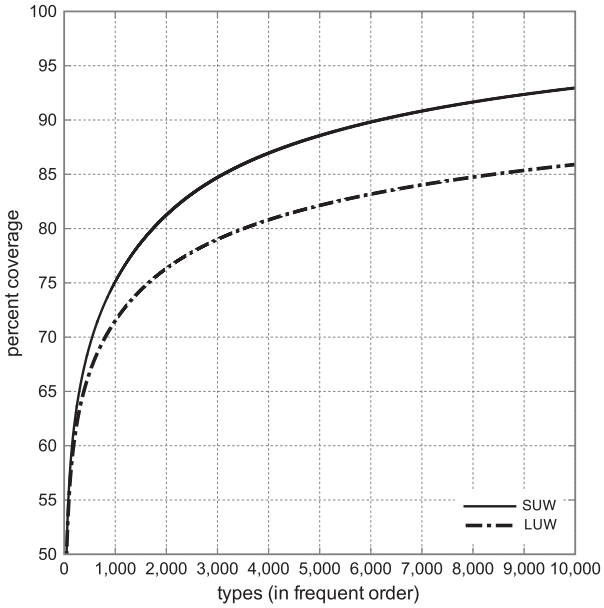

Figure 2: Type-coverage curves of Corpus-A

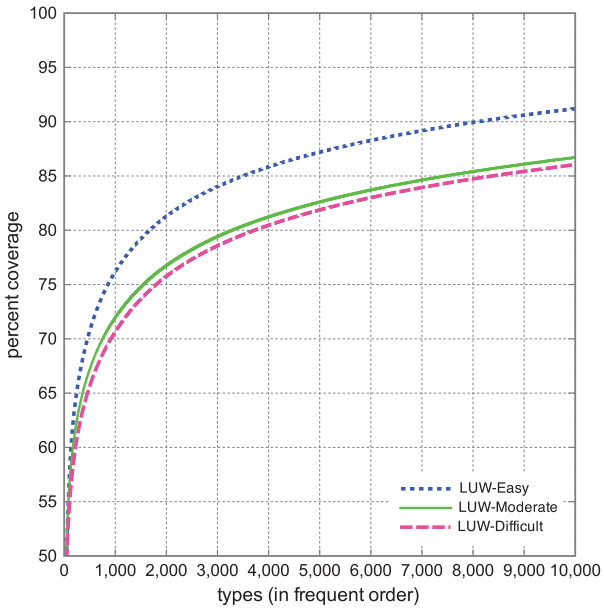

Figure 3: Type-coverage curves of LUWs in three readability bands

D-band exceeds that of the M-band at the top 319 frequent words, and that of the E-band at the top 879 frequent words. This unexpected result may be explained partially by the number of word types appeared in the D-band. The bottom of Table 2 shows the number of tokens and types of each bands in Corpus-B. Note that the text sizes of three bands of Corpus-B are nearly the same. From this table, we can see that the number of SUW types in the D-band is the smallest, and the average number of SUW types per sample in the D-band is also the smallest.

Why does a sample in the D-band have a relatively small number of SUW types? Our explanation to this question is the following.

- A sample in the D-band tends to have a dense content;

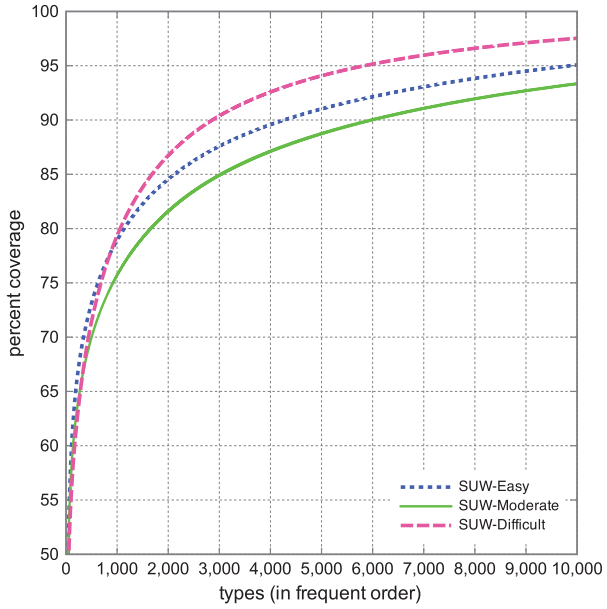

Figure 4: Type-coverage curves of SUWs in three readability bands

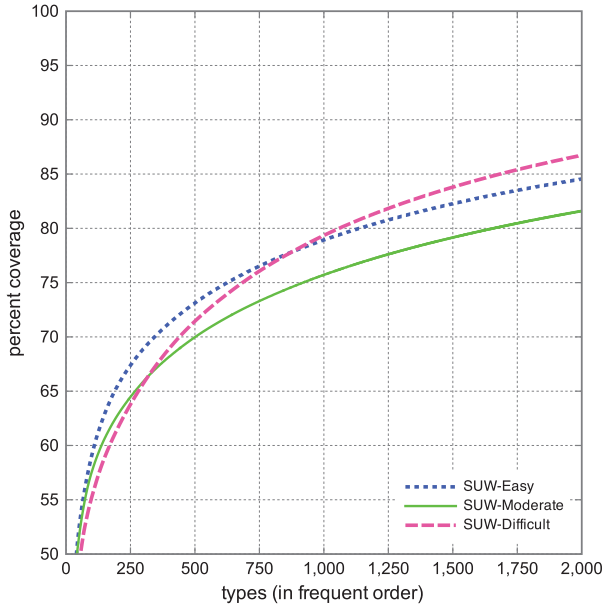

Figure 5: Type-coverage curves of the top 2,000 frequent SUWs in three readability bands

in other words, a sample focuses on a narrower topic. This reduces the number of SUW types appeared in a fixed-length sample.

- *Kango*, Chinese-origin words, are preferred to use in the D-band. From a kango, several semantically related compounds (LUWs) can be produced and these compounds tend to be used in a narrow topic. This also reduces the number of SUW types appeared in a fixed-length sample.

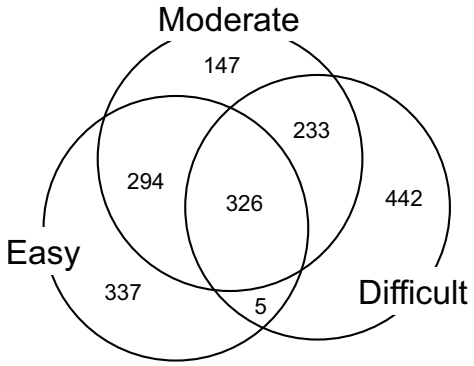

Figure 6: Venn diagram of the top 1,000 frequent SUWs in three bands

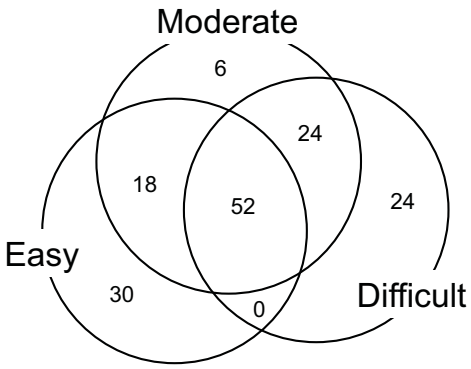

Figure 7: Venn diagram of the top 100 frequent SUWs in three bands

6. Common High-Frequent Words among Three Readability Bands

Another important finding in this study is that the number of common high-frequent words among different readability bands is smaller than we expected. We expected that most of high-frequent words are common in all kinds of books because they are functional or basic words.

Figure 6 shows the Venn diagram of the top 1,000 frequent SUWs of three readability bands. From this figure, we can see that the top 1,000 frequent SUWs of the E-band are quite different from that of the D-band. In fact, the intersection of these two lists is only 33% (=331/1000). Figure 7 shows the Venn diagram of the top 100 frequent SUWs of three readability bands. Even in the top 100 frequent SUWs, the intersection of the E-band and the D-band is 52%. These facts indicate that high-frequent words vary fairly according to readability bands.

It is well known that word frequency is the most important and reliable criterion to determine a basic vocabulary, while several shortcomings exist such as absence of certain important words and existence of certain difficult words in high-frequent word list (Nation, 1990). It is expected that the consideration of the readability of the target corpus would overcome these shortcomings and automate the compilation of a basic vocabulary (Brooke et al., 2012).

7. Acknowledgments

This study used the Balanced Corpus of Contemporary Japanese (BCCWJ) compiled by the National Institute for Japanese Language and Linguistics. This work was supported by JSPS KAKENHI Grant Number 24300052.

8. References

Julian Brooke, Vivian Tsang, David Jacob, Fraser Shein, and Graeme Hirst. 2012. Building readability lexicons with unannotated corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PITR '12, pages 33–39. Association for Computational Linguistics.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

I.S.P. Nation. 1990. *Teaching and Learning Vocabulary*. Heinle & Hinle Publishers.

Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yu Hara. 2011a. 『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版 (上). Technical Report LR-CCG-20-05-01, the National Institute of Japanese Language and Linguistics.

Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yu Hara. 2011b. 『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版 (下). Technical Report LR-CCG-20-05-02, the National Institute of Japanese Language and Linguistics.

Satoshi Sato and Wakako Kashino. 2012. Which text is easier? —judgment by human and machine— (in Japanese). In *Proc. of JCLWS-01*. National Institute for Japanese Language and Linguistics.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic assessment of Japanese text readability based on a textbook corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Satoshi Sato. 2011. Measuring text readability based on balanced corpus (in Japanese). *IPSJ Journal*, 52(4):1777–1789.

Norbert Schmitt. 2010. *Researching Vocabulary: a Vocabulary Research Manual*. Palgrave Macmillan.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting texts by readability. *Comput. Linguist.*, 36(2):203–227.