Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese

†Kikuo Maekawa, †Makoto Yamazaki, †Takehiko Maruyama, †Masaya Yamaguchi,

†Hideki Ogura, †Wakako Kashino, †Toshinobu Ogiso, ‡Hanae Koiso, and †‡Yasuharu Den

†Department of Corpus Studies, National Institute for Japanese Language and Linguistics
‡Department of Linguistic Theory and Structure, National Institute for Japanese Language and Linguistics
†‡Faculty of Letters, Chiba University

10-2, Midori-cho, Tachikawa-shi, Tokyo JAPAN 190-8561
E-mail: {kikuo, yamazaki, maruyama, masaya, ogura, waka, togiso, koiso}@ninjal.ac.jp, den@cogsci.l.chiba-u.ac.jp

Abstract

Compilation of a 100 million words balanced corpus called the *Balanced Corpus of Contemporary Written Japanese* (or BCCWJ) is underway at the National Institute for Japanese Language and Linguistics. The corpus covers a wide range of text genres including books, magazines, newspapers, governmental white papers, textbooks, minutes of the National Diet, internet text (bulletin board and blogs) and so forth, and when possible, samples are drawn from the rigidly defined statistical populations by means of random sampling. All texts are dually POS-analyzed based upon two different, but mutually related, definitions of 'word.' Currently, more than 90 million words have been sampled and XML annotated with respect to text-structure and lexical and character information. A preliminary linear discriminant analysis of text genres using the data of POS frequencies and sentence length revealed it was possible to classify the text genres with a correct identification rate of 88% as far as the samples of books, newspapers, whitepapers, and internet bulletin boards are concerned. When the samples of blogs were included in this data set, however, the identification rate went down to 68%, suggesting the considerable variance of the blog texts in terms of the textual register and style.

1. Introduction

One of serious problems in the corpus-based analysis of the present-day Japanese is the lack of a balanced corpus. Traditionally, most analyses are based upon three sources, namely, text archives of newspapers, a collection of copyright-expired literary works (*Aozora bunko*), and text obtained by internet crawling. Putting aside the problems of the copyright-expired texts, which are definitely too old to serve as material for the study of contemporary Japanese, the lack of a balanced corpus imposes two mutually related problems on linguistic studies.

For one thing, most of newspaper articles are written by journalists who are highly trained with respect to writing style. Accordingly, newspaper articles constitute a genre of Japanese text where linguistic variations of all sorts (orthographic, morphological, syntactic, and semantic) are suppressed to the minimum level. On the other hand, texts on the www are very much likely to include various registers and genres. It is also expected that a considerable amount of linguistic variation will be observed. It is, however, very difficult, if not impossible, to conduct analyses of style differences and/or linguistic variation using internet texts, because information about the genre of texts and/or the writers is usually missing. Moreover, the amount of retrieved texts can often be too large to be classified by hand.

To solve these problems in Japanese corpus linguistics, the National Institute for Japanese Language and Linguistics (NINJAL, hereafter) launched a corpus compilation project in the spring of 2006, aiming at the public release of Japan's first 100-million-word balanced corpus in 2011. The corpus is named the *Balanced Corpus of Contemporary Written Japanese*, or BCCWJ.

2. Design

BCCWJ consists of 3 component sub-corpora as shown in Fig.1. One of the most important characteristics of the BCCWJ is its sampling strategies to assure corpus representativeness: more than two-thirds of the whole corpus consists of samples drawn randomly from well-defined statistical populations. In Fig.1, the publication sub-corpus (PSC) and the library sub-corpus (LSC) consist exclusively of randomly selected samples. The population of the PSC is the whole body of books, magazines, and newspapers published during the years 2001-2005 (whose total size is estimated to be 65 billion characters and which, in turn, is estimated to contain about 38.5 billion words), and the population of the LSC is the books that are registered in more than 13 public libraries in the Tokyo metropolis and published after 1986 (47 billion characters and 27.5 billion words). The sizes of PSC and LSC after sampling are 35 million and 30 million words (in terms of SUW, see below) respectively.

PUBLICATION (PRODUCTION) SUB-CORPUS	**LIBRARY (CIRCULATION) SUB-CORPUS**
Books, Magazines, and Newspapers published during 2001-2005	Books published during 1986-2005
35 million words	30 million words
SPECIAL-PURPOSE SUB-CORPUS Whitepaper, Diet minutes, Web texts, Textbooks, etc., published during 1976-2005	
35 million words	

Fig.1 Components of the BCCWJ

The last sub-corpus of the BCCWJ is called 'special purpose' sub-corpus (SSC). This sub-corpus covers the text genres that are indispensable for the research projects in the NINJAL but are not covered either by the PSC or by the LSC: Web texts, school textbooks, governmental whitepapers, minutes of the National Diet, bestselling books, and so forth. The size of SSC is 35 million words. As for sample length, two texts of different length are taken from the same material (i.e., a book, a magazine, a newspaper, etc.) drawn from the population. One of them has a uniform length of 1,000 characters and is called a fixed-length sample. The other type, called a variable-length sample, has variable text length depending on the structure of the original text. A variable-length sample covers a well-defined and meaningful textual segment such as a section or a chapter. In the case of literary books and newspaper articles, the mean lengths of variable-length samples are about 4,000 and 1,000 characters, respectively.

3. Annotation

Sampled texts were annotated with respect to bibliographical information, character information, text structure, and POS information, and they are distributed as XML documents.

Tags about character information include "missingChar," "correction," and so on. The "missing Char" tag is applied when there is a character that is not included in the JIS0213:2004 character set, as in the upper panel of Fig.2. This tag is applied mostly to Chinese characters (*Kanji*). The "correction" tag, on the other hand, is applied when there is a typo. The character(s) causing the type (again, usually a Chinese character) is replaced with the seemingly correct one, but the original typo is retained as an attribute of the tag, as in the lower panel of Fig.2.

<missingChar attribute="HanIdeograph" unicode="U+5AEB" daikanwa="M06673" description="女偏に莫">＝</missingChar>

基<correction type="erratum" originalText="盟">盤</correction>投資

Fig.2 Tags about missing characters and corrections.

Tags about text structure include "sample", "article", "title", "cluster", "abstract", "figureBlock", "list", "paragraph", and "sentence". These tags are used to represent the hierarchical structure of the sample text. There are also tags about quotations and citations.

POS information is provided for two different levels of 'word'; short unit word (SUW) and long unit word (LUW). The dual POS analysis, which was first introduced in the *Corpus of Spontaneous Japanese* (Maekawa et al., 2000), was useful in the linguistic analysis of Japanese, a highly agglutinative language. Fig.3 compares the results of SUW and LUW POS analyses of /kokuricukokugokeNkyuHjo nioite wa/ (at the National research institute for Japanese language). The first 4 SUW nouns are reanalyzed as a single compound LUW noun. And, the three following SUWs (i.e., /ni oi te/) are reanalyzed as a single coupound particle in the LUW analysis.

A new SUW-based machine-readable dictionary, called UniDic, was developed for the POS analysis of BCCWJ and used with the *MeCab* morphological analyzer (Kudo, Yamamoto and Matsumoto 2004). See the next section for its performance.

4. The Status Quo

As of February 2010 (45 months since the beginning of the BCCWJ project), the total number of SUWs sampled so far is about 90 million. More than 90% of the sampled texts have been annotated with respect to bibliographical, character, and text structure information.

Fig.4 shows the current performance of automatic SUW analysis by the combination of UniDic and MeCab (see above). The bar denoted as "boundary only" stands for the accuracy (F-value) of the identification of SUW boundaries. "Boundary & POS" stands for the cases where POS information was correctly identified in addition to the boundary information. And, "boundary & POS & lexeme" stands for the cases where lexeme identity was correctly identified in the case of homonyms in addition to boundary and POS information.

The performance differed depending on the genre of text. The most difficult genre among written texts is Web texts, which are almost as difficult as the spontaneous spoken speech recorded in the *Corpus of Spontaneous Japanese* ("Spoken" in Fig.4). But even in the case of Web texts, the F-value of "Boundary & POS & Lexeme" is as high as 98%, which was the target value at the beginning of the project.

As for the clearance of copyright, which is probably the most difficult issue in the construction of present-day language corpora, texts corresponding to about 50 million words (SUW) have been copyright-cleared. For example, among the total of 24,050 book samples, we have so far made contact with the copyright holders of 19,971 samples (83.0%) as of the end of February 2010.

Japanese	国立	国語	研究	所	に	おい	て	は
Reading	kokuricu	kokugo	keNkyuH	Jo	Ni	oi	te	wa
GLOSS	national	language	research	institute	CASE	regarding	CASE	TOPIC
SUW	noun	noun	noun	suffix	Particle	verb	particle	particle
LUW	noun (compound)				particle (compound)			particle

Fig.3 Comparison of SUW and LUW.

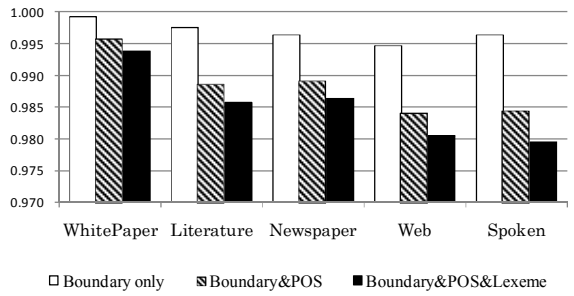

Fig. 4 Performanceof POS analysis.

TYPE OF RESPONSE	N
Permission	13,665
Denial	917
No response so far	4,424
No response at all	965

Table. 1 Responses of the contacted copyright holders (Case of book samples)

Table 1 shows the responses from the contacted copyright holders in the case of book samples. The ratio of permission is 68.4%, but this does not mean that over 30% of samples were denied by the copyright holders. Literal denial was less than 5%, as shown in Table 1. In this table, the rows labeled 'No response so far' and 'No response at all' both stand for cases where we did not receive a response from the copyright holders; the difference between them consists in the number of times we made contact with the copyright holders. In the case of 'so far' we have made contact only once, while in the case of 'at all' we have made contact more than one time. We will continue to try to obtain permission from the copyright holders, but at the end of the term of the project, we will be forced to make decision regarding the treatment of these 'no response' cases. If we take the liberty of interpreting the lack of response as a sign of implicit permission, we will be able to make more than 95% of the samples publicly available.

Copyright-cleared materials are available for full-text retrieval on the web (http://www.kotonoha.gr.jp/demo/) for the sake of demonstration. As of March 2010, about 46 million words are available at the above site.

5. Preliminary analyses

Evaluation of a balanced corpus can be done from various points of view, but the most important consideration should be the breadth of the textual variety of the corpus. A simple document classification task was conducted using a small subset of BCCWJ in order to evaluate the genre-related textual differences. The subset included 1,049,533 SUWs consisting of samples from white papers (W), internet bulletin boards (C), books (B), newspaper articles (N), and blogs (Y), as shown in Table 2. These samples were extracted randomly from the BCCWJ under construction. Relative distributions of SUW-POS categories were computed for each of the genres and are shown in Table 3. As can be seen from the table, the POS distribution was not uniform across sample genres; notable differences emerged in adverbs, auxiliary verbs, pronouns, adjectives, and suffixes.

To see the differences in a more comprehensible manner, linear discriminanat analysis was conducted. Two separate analyses were conducted using the data with and without the blog samples.

GENRE	N. File	N. SUW
White paper (W)	62	228,651
Bulletin board (C)	938	110,649
Books (B)	83	234,540
Newspaper (N)	340	360,814
Blog (Y)	497	114,879

Table 2. Subset data used in the preliminary analysis

POS	White paper (W)	Bulletin board (C)	Book (B)	News paper (N)	Blog (Y)
Particle	215	275	284	232	217
Adverb	4	18	18	6	15
Auxiliary verb	41	119	94	56	79
Verb	96	121	130	93	91
Noun	404	230	248	375	283
Pronoun	2	13	16	4	10
Adjective	5	20	15	8	14
Adjectival verb	11	10	11	8	9
Adnominal	5	6	10	4	5
Interjection	0	1	1	0	3
Conjunction	7	2	4	2	3
Prefix	8	6	5	8	8
Suffix	66	26	33	60	37
Marks	120	142	114	127	164
Others	18	10	16	18	62

Table 3. Distribution (per thousand) of SUW-POS categories in the data

GENRE	Mean length	SD
White paper (W)	42.5	9.2
Bulletin board (C)	17.6	6.1
Books (B)	26.6	7.4
Newspaper (N)	24.1	5.1
Blog (Y)	16.5	13.5

Table 4. Mean and SD of sentence length.

The data for the LDA involved the POS distribution data (Table 3) and mean sentence length data (in terms of the number of SUWs in a sentence, Table 4) of genres W, C, B, N, and Y. The results of leave-one-out cross validation are shown in Tables 5 and 6, whose rows and columns stand respectively for observations and predictions; the cells on the diagonal show the numbers of correctly identified

samples. The overall correct discrimination rate was about 88.3% in Table 5, where blog samples (Y) were excluded from the analysis. Samples of bulletin boards (C), newspapers (N), and whitepapers (W) were classified correctly, while most of the book samples (B) were missclassified as bulletin board samples (C). Table 6 shows that addition of blog data brings the correct identification rate down to 69.2%. Most of the blog samples (Y) were missclassified as bulletin board samples.

	B	C	N	W
B	22	47	8	6
C	16	873	45	4
N	2	29	307	2
W	0	0	7	55

Table 5. Result of cross-validation. Analysis without the blog samples.

	B	C	N	W	Y
B	7	65	7	4	0
C	7	830	22	3	76
N	0	40	278	2	20
W	0	0	13	49	0
Y	3	223	88	18	165

Table 6. Result of cross-validation using the all data.

Lastly, the distribution of samples on the planes defined by the first two linear discriminant functions (LD1 and LD2) are shown in Figures 6 and 5, which show respectively the results with and without the blog samples. Note that these results were obtained by analyses that are independent from the ones conducted for Tables 5 and 6: the LDA for Figs. 5 and 6 were not leave-one-out cross validation.

In both of these figures, the contribution of the abscissa (LD1) is considerable. LD1 separates the data clouds corresponding to the samples of whitepaper (W), newspapers (N) and bulletin boards (C) with considerable accuracy. As opposed to this, the contribution of the ordinate (LD2) is much less clear. It seems that the contribution of LD2 is limited to the separation between the clouds of W and N. As a matter of fact, the proportion of traces of LD1 and LD2 were 0.855 and 0.130 in the case of Fig. 5, and, 0.716 and 0.173 in the case of Fig. 6. Closer look at the coefficients of LD1 reveals that the most important factors for LD1 involve sentence length (sentences are shorter at the upper edge of LD1), relative frequency of nouns (abundant at the upper edge), and relative frequency of auxiliary verbs (abundant at the upper edge).

Comparison of Figs. 5 and 6 reveals the special nature of the blog data. The basic structure of the LD1-LD2 plane seems to be the same in these figures. The main difference between the figures consists in the range of distribution of blog samples. In Fig. 6, the distribution of the blog samples (Y) covers virtually the whole range of both LD1 and LD2. This diversity of the blog sample was reflected in the STD value in Table 4 above.

To conclude, users of the BCCWJ are able to have easy and secured (in terms of copyright violation) access to much wider range of texts than that of the newspaper archives that have been the main resource of corpus linguistics involving the Japanese language. BCCWJ will be publicly available in the year of 2011 as scheduled.

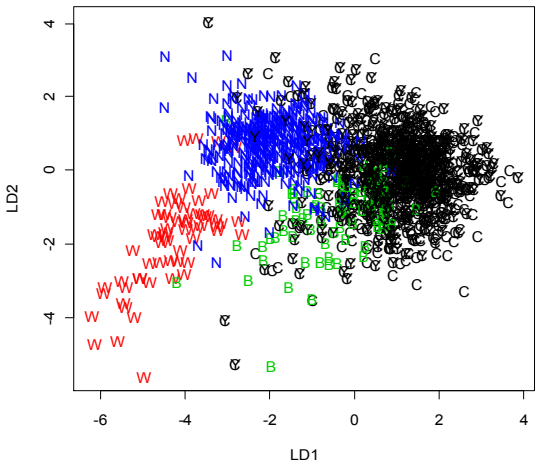

Fig. 5 Distribution of samples on the LD1-LD2 plane. Data excluding the blog samples. See Table 4 for plotting symbols.

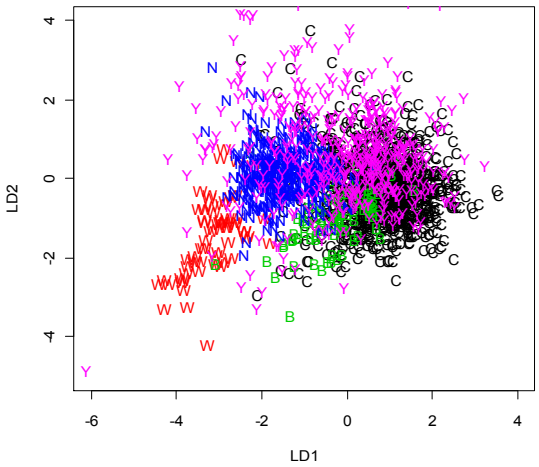

Fig. 6 Distribution of samples on the LD1-LD2 plane. The whole samples. See Table 4 for plotting symbols.

References

Kudo, T., K. Yamamoto, and Y. Matsumoto (2004). "Applying Conditional Random Fields to Japanese Morphological Analysis," *Proc. EMNLP 2004*, pp. 230-237.

Maekawa, K., H. Koiso, S. Furui, and H. Isahara (2000). "Spontaneous speech corpus of Japanese," *Proc. 2nd LREC*, pp.947-952.