Lang Resources & Evaluation (2014) 48:345-371
DOI 10.1007/s10579-013-9261-0

ORIGINAL PAPER

Balanced corpus of contemporary written Japanese

Kikuo Maekawa · Makoto Yamazaki · Toshinobu Ogiso · Takehiko Maruyama · Hideki Ogura · Wakako Kashino · Hanae Koiso · Masaya Yamaguchi · Makiro Tanaka · Yasuharu Den

Published online: 29 December 2013

Abstract The balanced corpus of contemporary written Japanese (BCCWJ) is Japan's first 100 million words balanced corpus. It consists of three subcorpora (publication subcorpus, library subcorpus, and special-purpose subcorpus) and covers a wide range of text registers including books in general, magazines, newspapers, governmental white papers, best-selling books, an internet bulletin-board, a blog, school textbooks, minutes of the national diet, publicity newsletters of local governments, laws, and poetry verses. A random sampling technique is utilized whenever possible in order to maximize the representativeness of the corpus. The corpus is annotated in terms of dual POS analysis, document structure, and bibliographical information. The BCCWJ is currently accessible in three different ways including *Chunagon* a web-based interface to the dual POS analysis data. Lastly, results of some pilot evaluation of the corpus with respect to the textual diversity are reported. The analyses include POS distribution, word-class distribution, entropy of orthography, sentence length, and variation of the adjective predicate. High textual diversity is observed in all these analyses.

K. Maekawa (✉) · M. Yamazaki · T. Ogiso · T. Maruyama · W. Kashino · M. Yamaguchi · M. Tanaka
Department of Corpus Studies, National Institute for Japanese Language and Linguistics (NINJAL), Tokyo, Japan
e-mail: kikuo@ninjal.ac.jp

H. Ogura
College of Letters, Ritsumeikan University, Kusatsu, Japan

H. Koiso
Department of Linguistic Theory and Structure, NINJAL, Tokyo, Japan

Y. Den
Faculty of Letters, Chiba University, Chiba, Japan

Keywords BCCWJ · Japanese · Balanced corpus · Design · Annotation · Dual POS analysis · Evaluation · *Shonagon* · *Chunagon*

1 Introduction

One serious problem in the corpus-based analyses of present-day Japanese is the lack of a balanced corpus. To solve this problem, the authors have recently released a 100 million words balanced corpus of contemporary written Japanese. The aim of this paper is twofold: to describe the basic properties of the corpus and to evaluate the corpus from a point of view of the diversity of texts in the corpus. In this introductory section, the reason we need a balanced corpus in the study of Japanese is discussed.

Most corpus-based analyses of the Japanese language published in the last two decades or so depends heavily upon the analyses of the text archives of newspaper articles released by newspaper companies. For example, the Kyoto University Text Corpus (Kurohashi and Nagao 1998) that played an important role in the development of an annotated corpus of Japanese natural language processing (NLP) consists of 40 thousands sentences taken from the articles of the *Mainichi* newspaper published in 1995.

Use of newspaper articles in linguistics and natural language processing, however, imposes several important problems with respect to the corpus representativeness, because newspaper articles are written by skilled journalists and independently checked by professional proofreaders. Accordingly, as we will see later in Sect. 6, newspaper articles belong to the class of Japanese text where linguistic variations are considerably suppressed.

In addition to newspaper articles, literary texts (mostly novels and essays) are also used in the linguistic analyses of Japanese. Two main sources in this field include the CD-ROM version of *Shincho bunko no hyakusatsu* (one hundred titles from the Shincho paperbacks), and *Aozora bunko,* a collection of copyright-expired literary works.[1] The problem with these collections is that the texts are too old to be called ‘contemporary.’

In addition to these, texts gathered by internet crawling have become a major resource in recent years for both linguistics and NLP. In linguistics, texts on the web are expected to be suitable for the study of linguistic variation, because materials of diverse writing styles are likely to exist on the web. There is, however, a problem with this approach; so called meta-information about the texts (genre, registers, etc.) and/or the writers (gender, age, etc.) are frequently missing.

Although there is the possibility of estimating the meta-information by means of various up-to-date statistical clustering and classification methods, this approach requires a certain amount of supervised learning training data, derived from reliable reference corpora including the types of data mentioned above and covering various text types.

[1] http://www.aozora.gr.jp/.

To solve these problems, the authors launched a corpus compilation project in the spring of 2006, for public release of Japan's first 100 million words balanced corpus in the year of 2011. The corpus was named the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ, hereafter).

2 Corpus design

There are three basic principles in the design of the BCCWJ. The first principle is to design the corpus exclusively as a corpus of written Japanese rather than spoken Japanese. Spoken varieties are excluded from the corpus mainly because some of them are covered by the recently released *Corpus of Spontaneous Japanese*, or CSJ (Maekawa et al. 2000; Maekawa 2003). In the actual BCCWJ, however, there is one exception to this principle: The OM register (see below for abbreviation) that deals with the minutes of Japan's National Diet. The reason for the inclusion of this register is discussed in Sect. 2.2.3 below.

The second principle is the maximum use of random sampling. As is well known, samples extracted by this method maximally represent the characteristics of the population. Unfortunately, however, several reasons prevent the application of random sampling in the implementation of a balanced corpus. For one thing, application of random sampling is possible only when the population for the sampling is explicitly definable. As far as written Japanese is concerned, it was in most cases possible to get the information necessary for defining the population. In Sect. 2.2 below, characteristics of each register are explained with reference to its sampling population.

The third principle is to make the corpus publicly available. This principle, when combined with the second one, inevitably increases the burden of copyright treatment. This problem is discussed in Sect. 3 below.

2.1 Size

The size of the BCCWJ is 100 million words. It is the size of renowned British National Corpus (BNC), but this is not the only reason with which the size of BCCWJ was determined. A balanced pilot corpus of one million words covering books, magazines, and newspapers was constructed in 2005 in order to estimate the construction cost of a large-scale balanced corpus. Based on this study, it turned out that one hundred million words is the maximum size, given a corpus design described in Sects. 2 and 4 below, and, more importantly, to be completed within 5 years. It was estimated that construction of a corpus larger than this size would certainly exceed the limit of the research budget available at that time.

2.2 Balance and representativeness

BCCWJ consists of three subcorpora: Publication, library, and special-purpose subcorpora. Also, the subcorpora can be divided into 13 different classes of texts

Table 1 Structure of the BCCWJ

Subcorpus	Register	Abbr.	# Sample	# Word (Unit: million)
Publication	Books	PB	10,177	28.55
	Magazines	PM	1,996	4.44
	Newspapers	PN	1,473	1.37
Library	Books	LB	10,551	30.38
Special-purpose	White papers	OW	1,500	4.88
	Bulletin board (*Yahoo! Chiebukuro*)	OC	91,455	10.26
	Blog (*Yahoo! Blog*)	OY	52,680	10.19
	Best-selling books	OB	1,390	3.74
	School textbooks	OT	412	0.93
	Minutes of the National diet	OM	159	5.10
	Publicity newsletters of local governments	OP	354	3.76
	Laws	OL	346	1.08
	Poetry verses	OV	252	0.25

that we tentatively call text 'registers.' Table 1 shows the relationship between the three subcorpora and 13 registers. The following subsections will be devoted to the description of these subcorpora and registers.

2.2.1 Publication subcorpus

The publication subcorpus, or PSC, consists of three registers, books in general (abbr. PB), magazines in general (PM), and newspapers from around Japan (PN). Samples of books and magazines are the most important among these registers. Texts in these registers are not readily available for linguists and NLP researchers except for the copyright-expired ones (see Sect. 1 above), because copyright clearance in these registers are extremely complicated, hence difficult (see Sect. 3 below).

From a point of view of corpus balance, the unique characteristic of the PSC is that a single definition of statistical population is shared by three component registers. The statistical populations for the sampling consisted of all books, magazines, and newspapers published in the years 2001–2005; the populations were constructed based on the information provided by publicly available databases including J-BISC (Japan Biblio-Disc, a book directory service of the National Diet Library) and *Zasshishinbunsokatarogu* (almanac of all magazines and newspapers).

As far as the PSC is concerned, accordingly, the mutual ratio of the sample sizes of the three registers (i.e. 28.55 vs 4.44 vs 1.37) is not arbitrary. They rather reflect the activities of the three registers in the designated years, as measured by the amount of published texts.

In the actual sampling of the samples belonging to the PSC, the technique known as stratified (or layered) sampling is used. For example, the population of the book samples is divided into 55 categories covering all 11 categories of the NDC (Nippon

Decimal Classification, Japan's standard book classification system) for each of the 5 years, so that samples are extracted from all these categories. In the same vein, the population of the magazine samples is divided into 30 categories (6 magazine genres and five years), and, the population of newspaper articles is divided into 80 categories (16 newspapers and five years).

The populations thus defined contain 48.5, 10.5, and 6.4 billion characters respectively for PB, PM, and PN registers. The sampling ratio of PSC is set to one thousandth (1/1,000). Note that the size of population is estimated in terms of the number of characters rather than words because Japanese texts need word segmentation before they are counted as words.

Lastly, it is important to understand that the PSC is constructed exclusively on the basis of the production (publishing) aspect of publication and has nothing to do with the reception (sales) aspect. Two books in the same format and having the same number of pages always have the same probability of being extracted as a sample of the PSC regardless of the difference in the sales records of the books.

2.2.2 Library subcorpus

Library subcorpus, or LSC, consists of a single register of LB, and is designed so that it reflects a reception aspect of publication. Needless to say, the best way to achieve this goal is to construct a corpus whose materials are sampled on the basis of the sales data, but such a corpus is impossible to design, because reliable data does not exist regarding the sale of books and/or magazines.

Instead, a corpus is designed whose population is all the books registered in the public libraries of Tokyo Metropolis, with the expectation that the books registered in multiple public libraries which have stayed there for a while may represent, in some way, the books read (namely, received) by a certain amount of readers.

More precisely, the statistical population of the LSC is the set of books satisfying the following two conditions. First, the book is published between 1986 (the year when the ISBN, which is indispensable for the management of library databases, was adopted by most Japanese publishers) and 2005. Second, the book is registered in the public libraries encompassing more than 13 cities and/or special wards (out of the total of 52) of Tokyo metropolis. The size of the population thus defined (47.9 billion characters) is nearly the same as that of the population of book samples in the PSC. The LSC population is divided into 220 layers consisting of 11 NDC categories and 20 years of publication for stratified sampling.

2.2.3 Special-purpose subcorpus

The special-purpose subcorpus, or SSC, is designed so that it covers the kinds of registers that are regarded to be indispensable for the understanding of the totality of contemporary written Japanese, but are not covered either by PSC or LSC. As shown in Table 1, there are nine registers in the SSC.

White paper (OW) is the collection of samples randomly sampled from the population of 1,006 white papers published by the Japanese government during the years 1976–2005. It is expected to represent the variety of Japanese used in official administrative publications. The population for the OW samples is divided into 54 layers (consisting of nine genres and six time periods of five years).

Bulletin-board (OC) and blog (OY) are the two registers representing the texts in the internet. They are expected to represent the most up-to-date characteristics of the Japanese language. At the same time, they are expected to represent the writing style that is more casual than the other registers. Moreover, the OC samples are expected to show more interpersonal features of the language like phrase-final particles and honorifics than in the other registers; this expectation is based on the fact that exchanging a series of questions and answers in a bulletin-board is all written dialogues.

The population for the OC samples consists of more than three million pairs of questions and the corresponding answers (including multiple answers to a single question) posted between October 2004 and October 2005. The population for the OY samples consists of about 3.5 million blog articles posted between April 2008 and April 2009.

Best-selling books (OB) contain samples from the population consisting of 951 best-selling books published during the years 1976–2005.

School textbooks (OT) consist of samples from the population consisting of 145 elementary, junior-high, and high school textbooks published during the years 2005–2007. The population was layered with respect to nine subjects and three types of school.

Minutes of the National Diet (OM) consist of samples from the population consisting of 32,986 diet meetings ranging between 1976 and 2005. The population was layered with respect to the type of diet (the House of Councilors and the House of Representatives), six time periods of five years each, and the type of meetings.

As pointed out earlier, OM is the sole exception to the first principle of BCCWJ design (exclusion of spoken varieties). The inclusion of OM is thought to be necessary because, for one thing, it is concerned with the language of political discussions and debates that is not covered by other registers of the BCCWJ and the CSJ. For another, there is a need of the corpus users for the inclusion of this register. Studies collected in Matsuda (2008) show convincingly the usefulness of the minutes of the National Diet for the study of language variations and changes.

Publicity newsletters of local governments (OP) consist of samples from the population of newsletters issued by 100 local municipalities in the year of 2008. The municipalities covers all 8 areas of Japan (Hokkaido, Tohoku, Kanto, Chubu, Kinki, Chugoku, Shikoku, and, Kyushu-Okinawa), and the population was layered along these areas.

Laws (OL) consist of samples from the population consisting of 718 laws that were promulgated during the years 1976–2005. The population was layered for six time periods of five years each.

Lastly, poetry verse (OV) consists of samples of short fixed-form verses (tanka and haiku) and free-form poems randomly selected from the following three collections: Vols. 14–17 of Chikuma Shobo’s *Gendai Tanka Zenshu* (2002), vols. 8–15 of Kadokawa’s *Zoho Gendai Haiku Taikei* (1980–82), and 118 titles of Shichosha’s *Gendai Shi Bunko* (1986–2005). Note that in register OV, multiple tanka and haiku are included in a single sample (see Sect. 2.3.3 below).

2.3 Notes on corpus design and implementation

In this section, some additional information about the design and implementation of the BCCWJ are presented. This information is indispensable for the users of the corpus.

2.3.1 Missing registers

There are two registers that the present authors wanted to include in the BCCWJ but to no avail: internet mail and manga (Japanese cartoons). Internet mail often contains materials too private to be publicly available. As for manga, the main obstacles consist in the difficulties of copyright treatment and estimation of the amount of text in a title.

2.3.2 Temporal coverage

As described in the preceding subsections, the temporal coverage of a population differs considerably depending on the registers. Registers OW, OB, OM, and OL have the longest coverage of 30 years (1976–2005). LB (LSC as a whole) has a middle coverage of 20 years (1986–2005). PB, PM, and PN (PSC as a whole) have short coverages of five years (2001–2005). The coverage of OT is also short (2005–2007), but not identical to that of PSC. OC, OY, and OP have the shortest coverage of one year.

Users have to be careful about the difference in the temporal coverage when he/she wants to compare the results obtained from different registers. On the other hand, it is possible to obtain knowledge about the real-time change of the Japanese language in the past 20–30 years using the registers of relatively long temporal coverage like LB, OB, OW, OM and OL.

2.3.3 Sample length

Sample length is an important factor of sampling. From the point of view of corpus representativeness, it is desirable to extract from the population many very short samples of the same length rather than one extremely long sample, given that the total amount of words in the samples stays the same. However, extraction of very short samples is not necessarily the best solution from a point of view of linguistics research. For example, too short a sample makes it impossible to disambiguate homonyms. It also makes it impossible to analyze the structure of discourse. Those who are interested in the structure of discourse may want to have access to as long a context as possible.

As a remedy to this problem, two different sample lengths are adopted in the sampling of the BCCWJ: fixed-length and variable-length samples. A fixed-length sample consists of 1,000 characters (counting kanji—Chinese logographs—, hiragana, katakana—two types of Japanese syllable letters–, and Roman alphabets letters equally as one character, and excluding all punctuation marks). Fixed-length samples are suitable for various statistical inferences like the estimation of word and/or kanji frequencies. A sample of 1,000 characters roughly approximates the characters printed in two pages of a Japanese paperback (文庫本). Needless to say, the end of a fixed-length sample does not always match a linguistic boundary of any sort (like phrase and sentence). When the thousandth character of a fixed-length sample occurs in the middle of a sentence, the end of the sample is extended to match up with the nearest sentence boundary.[2]

A variable-length sample covers the entire text which is logically organized as a chapter or section. The length of a variable-length sample differs considerably depending on samples. For example, some newspaper articles like obituaries are shorter than 1,000 characters. On the other hand, there are novel chapters that encompass several tens of thousands characters. In the case of the latter, we introduce an upper limit of the variable-length samples; a single variable-length sample may not exceed 10,000 characters. Variable-length samples are suitable for various context-sensitive linguistic analyses.

Notice that variable-length samples are extracted from all BCCWJ registers, while fixed-length samples are extracted only from PB, PM, PN, LB, and OW. In these registers, there is usually overlap between the fixed- and variable-length samples, because the two samples are extracted simultaneously from the same (randomly selected) document. When we compute the word frequency of the BCCWJ as in Table 1, the overlap was excluded from the data. See also the description of *Chunagon* in Sect. 5.2 below.

3 Treatment of copyright

In present-day society, one of the largest difficulties of corpus compilation lies in the treatment of copyright. It is especially true of a corpus like BCCWJ that exclusively deals with contemporary materials. All BCCWJ samples are copyright-protected, with the following two exceptions. Texts of law (register OL) are not copyright-protected according to the Japanese copyright law. The minutes of the National Diet (OM) are copyright protected, but the councilors and representatives make it a rule to abandon their rights.

There are two special factors that make the copyright treatment of the BCCWJ very difficult. The total number of samples in the BCCWJ is 172,675 (counting only variable-length samples), and is much larger than in BNC (which consists of about

[2] The location of the thousandth character is marked by a tag so that users can extract exactly 1,000 characters when it is needed.

4,000 samples). This is because the sample-length of the BCCWJ is generally much shorter than that of the BNC. This is true not only in the fixed-length samples, but also in the case of variable-length samples.

Moreover, extensive use of random sampling makes the copyright treatment even more difficult. Because the choice of samples is random, hence has nothing to do with the easiness of copyright treatment, it is impossible to give priority to the samples whose authors are likely to give permission to the use of their samples in the corpus. A brief overview of the copyright treatment in the construction of the BCCWJ is presented below.

Samples in registers OC and OY are the easiest to clear copyright. Yahoo! Japan Corporation, who holds the copyright of all materials posted to their internet site, gave us permission to all samples. Similarly, the permissions to the samples in the PN and OW registers are given by the newspaper companies and relevant ministries or agencies in the Japanese government.

Samples in other registers are treated basically on the basis of one-by-one negotiation. In the case of book samples in registers PB, LB, and OB, for example, we obtained permissions from the authors for more than 24 thousands samples (including many cases where an author has multiple samples). Needless to say, the negotiations are not always successful. The success rate is about 80 %, once we are successful in making contact with the authors (or copyright holders). Because the probability of successful contact with the authors is about 90 %, the final success rate of the negotiation is about 70 %. It was hence necessary to prepare more samples than we really need. In the case of book samples in the PB, LB, and OB registers, for example, the total number of prepared samples is 29,188 while the number of samples currently in the BCCWJ is 22,058.

4 Corpus annotation

At the present time, three basic annotations are applied to the BCCWJ, namely, POS annotation, document structure annotation, and meta-information annotation.

4.1 POS annotation

4.1.1 Dual POS analysis and the UniDic

As suggested earlier in Sect. 2.2.1, POS analysis of Japanese text requires word segmentation. It hence requires principled treatment of a linguistic unit approximating a ‘word’. In this respect, however, treatment of words, compound words especially, in the preceding POS analysis systems is very problematic. Table 2 compares the outputs of three existing machine-readable dictionaries of automatic POS analysis systems, viz., Juman, Yahoo! Japan’s API, and Chasen with IPA dictionary. The output of MeCab with UniDic, a newly developed dictionary for the POS annotation of the BCCWJ (Den et al. 2008; See also the next subsection) is also shown for further comparison. The table shows word segmentation done

Table 2 Comparison of existing dictionaries

DICTIONARY	NAMES OF NATIONAL INSTITUTIONS			
	国立国会図書館 National Diet Library	国立公文書館 National Archives of Japan	国立民族学博物館 National Museum of Ethnology	国立天文台 National Astronomical Observatory
Juman	国立+国会+図書+館	国立+公文書+館	国立+民族+学+博物+館	国立天文台
Yahoo!	国立国会図書館	国立+公文書館	国立民族学博物館	国立+天文台
IPA	国立+国会図書館	国立+公文書+館	国立+民族学博物館	国立天文台
UniDic	国立+国会+図書+館	国立+公+文書+館	国立+民族+学+博物+館	国立+天文+台

automatically by the systems when the names of four Japanese national institutes are analyzed out of context. The '+' symbols show the locations of word boundaries.

国立国会図書館 ("kokuritsukokkaitoshokan" 'National Diet Library'; "kokuritsu" 国立 'national', "kokkai" 国会 'diet', and "toshokan" 図書館 'library') is treated in three different ways. The dictionary used in Yahoo! Japan's web API treats it as a single entry.[3] IPA (Information-technology Promotion Agency)'s dictionary delivered with the NAIST (Nara Institute of Science and Tchnology)'s ChaSen morphological analysis system (Asahara and Matsumoto 2000) divides it into two entries, viz., "kokuritsu" + "kokkaitoshokan". And, the dictionary of Kyoto University's JUMAN morphological analysis system (Kurohashi et al. 1994) divides it into four entries as "kokuritsu" + "kokkai" + "tosho" + "kan" (where "kan" 館 is analyzed as a suffix standing for 'house' and "tosho" 図書 stands for 'books').

What is worse, from a point of view of linguistics, considerable internal inconsistency is observed in all dictionaries except for the UniDic. In the case of Juman, the string 国立 "kokuritsu" is analyzed as a word in the first three institutions, but 'National Astronomical Observatory' makes an exception. The output of IPA shows exactly the same pattern. On the other hand, Yahoo! treats 'National Diet Library' and 'National Museum of Ethnology' as single entries, while it analyzes 'National Archives of Japan' and 'National Astronomical Observatory' into two entries. Similar within-dictionary inconsistency can be observed in the treatment of suffix 館 "kan" in the IPA. These internal inconsistencies can be a serious obstacle when researchers want to obtain basic statistical information about the structure of the language.

These inconsistencies seem to emerge as a result of the interaction of two factors. In the first place, Japanese is a so-called agglutinative language that has a high degree of freedom in the definition of 'word'. In addition to this, it seems that users of a morphological analysis system have heterogeneous requirements on the system's output. Users who are interested in the analysis of named entities may prefer single entry analysis, while the users who are interested in the morphological structure of compound nouns may prefer fine segmentation. Clearly, these demands

[3] http://developer.yahoo.co.jp/webapi/jlp/ma/v1/parse.html.

Table 3 Example of two-way POS analysis

SUW			LUW		
ENTRY	**POS**	**GLOSS**	**ENTRY**	**POS**	**GLOSS**
公害	Noun	pollution	公害紛争処理法	Noun	act for the settlement of pollution dispute
紛争	Noun	dispute			
処理	Noun	settlement			
法	Noun	act			
に	Particle	DATIVE	における	Particle	LOCATIVE
おけ	Verb	locate			
る	Aux. verb	PRESENT			
公害	Noun	pollution	公害紛争処理	Noun	settlement of pollution dispute
紛争	Noun	dispute			
処理	Noun	settlement			
の	Particle	GENITIVE	の	Particle	GENITIVE
手続	Noun	procedure	手続	Noun	procedure
は	Particle	TOPIC	は	Particle	TOPIC

are mutually incompatible, and the mixture of the two demands results in the lack of morphological consistency as observed in Table 2.

This problem can be resolved, at least partially, if the text is analyzed consistently at two different levels of morphology. One of the levels is called SUW, or short-unit word. SUW approximates the level at which entries of traditional Japanese dictionaries are specified. The other level is called LUW, or long-unit word. LUW is devoted mostly, but not exclusively, for compound words like compound nouns, compound verbs, and compound particles.[4]

Table 3 shows an example of the dual POS analysis. The example phrase is 公害紛争処理法における公害紛争処理の手続きは (“kogaifunsoshoriho niokeru kogaifunsoshori no tetsuzuki wa”, ‘as for the procedure of the settlement of pollution disputes according to the act for the settlement of pollution disputes’). The left and right halves of the table show the results of SUW and LUW analyses respectively. The first four nouns of SUW correspond to a compound noun in the LUW analysis, the following three SUW entries, consisting of a particle, a verb, and an auxiliary verb, correspond to a single compound particle in the LUW analysis, and the next three SUW nouns correspond to a compound LUW noun.

There are two more important facts to be pointed out in the table. First, words in LUW are not always compound words. As in the last three rows of the table, SUW and LUW analyses coincide if a text does not include any compound word.

Second, it can be the case that substrings of an LUW can be used as independent LUWs as is the case in the first and third LUW entries in Table 3, where the third entry

[4] The first corpus to which the dual POS analysis was applied was the Corpus of Spontaneous Japanese (Maekawa et al. 2000).

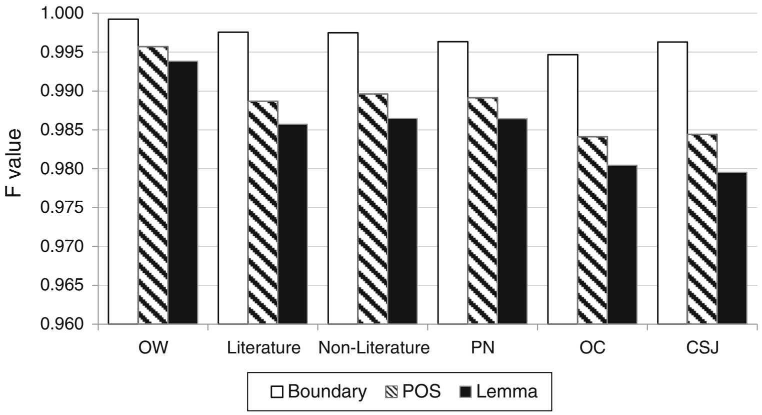

Fig. 1 Performance of SUW analysis by MeCab + UniDic

is the substring of the first. One more example is presented: 研究所 ("kenkyujo", 'research institution') is an LUW consisting of two SUWs, 研究 ("kenkyu", 'research') and 所 ("jo", 'institution'), but longer LUWs containing this LUW are also possible like 国語研究所 ("kokugokenkyujo", 'research institute of the Japanese language') or 国立国語研究所 ("kokuritsukokugokenkyujo", 'national institute of the Japanese language').[5]

In the POS analysis of the BCCWJ, texts are first analyzed into SUW using the UniDic; and then, some of the SUW are merged into LUW on the basis of statistical learning (Uchimoto and Isahara 2007; Kozawa et al. 2011). Although far from being perfect, introduction of the dual POS analysis resolves considerably the problems caused by the contradicting demands on POS analysis. The number of words is counted consistently in SUW in this paper.

4.1.2 The performance

The SUW analysis of the BCCWJ is conducted using the combination of the MeCab morphological analyzer (Kudo, Yamamoto, and Matsumoto 2004) and the UniDic. Figure 1 shows the performance of SUW analysis by MeCab and UniDic. In addition to the three registers of the BCCWJ (OW, PN, and OC), samples of books in PB, LB, OB registers are reclassified into the two categories of "literature" and "non-literature" here. Moreover, performance of the spoken data in the CSJ is also shown.

In this figure, the performance is evaluated by means of the F-measure, which is computed using the formula 2PR/(P + R), where P and R stand respectively for precision and recall. Samples used for the evaluation (about 100,000 SUW in size) are randomly selected from the BCCWJ-Core and the CSJ. They are analyzed manually and cross-checked by a group of expert annotators. The result of the

[5] The last entry is the Japanese name of the institution to which some of the authors belong, i.e., NINJAL.

manual analysis thus obtained is treated as the ‘correct answer’ in the following evaluations.

F-measure is computed under three different criteria of evaluation. Under the first criterion, only the correctness of SUW boundary locations is evaluated. The F-value was computed based upon the recall and precision values obtained by comparing the system output with the correct answer. The result is shown by the open bars in the figure.

Under the second criterion, conjoint correctness of the SUW boundary locations and the POS information is evaluated. In Fig. 1, the shaded bars annotated as ‘POS’ stand for this case. Lastly, under the third criterion, the conjoined correctness of boundary location, POS information, and lemma specification is evaluated. Here, lemma specification means the correctness of lemma choice from a set of synonyms. This criterion is important because Japanese has a considerable number of homonyms for historical and phonological reasons. For example, a word バス “basu” has at least five meanings: ‘bus’, ‘bass’, ‘bath’ (loans from English), ‘Bass’ (surname or place name in English), and ‘lotus’ (native Japanese “hasu” turns into “basu” by the rule of sequential voicing, or rendaku, when the word appears in the last half of a compound noun like in “onibasu” ‘Euryale ferox’). The filled bars in the figure stand for this case. Note also that the BCCWJ samples analyzed here are open data, meaning that they were not included in the data used for the learning of the SUW analysis system.

As can be seen from the figure, the system performance differs depending on registers, and the analysis of the web text (OC) is the most difficult among the BCCWJ registers. But the F-measures exceed the level of 0.98 even in the most difficult register and under the most difficult (i.e., the third) criterion. The latest version of UniDic is downloadable from the internet.[6]

As for the LUW analysis, Fujiike et al. (2010) reports the accuracies (but not the F-measure) computed under the same three criteria as in SUW, and across nearly the same registers as in Fig. 1. The reported accuracy is higher than 98 % even in the most difficult case, i.e. the analysis of texts in the OC register as evaluated according to the third criterion.

4.1.3 The BCCWJ-core

In the course of the SUW analysis of the BCCWJ, about 1 % of the corpus is analyzed with special care. This subset is named BCCWJ-Core. As far as the texts in this subset are concerned, the results of the automatic SUW analysis are checked and corrected manually so that the average accuracy becomes higher than 99 % even when they are evaluated with the third criterion mentioned above. The high-accuracy POS data thus obtained was utilized as the learning data for the training of the MeCab + UniDic analysis system. Also, the BCCWJ-Core should be quite useful for the corpus users who put more priority on the accuracy of morphological annotation rather than the amount of data. Table 4 shows the contents of the BCCWJ-Core.

[6] http://sourceforge.jp/projects/unidic/.

Table 4 Contents of the BCCWJ-Core

Register	# Sample	# Word
PB	83	204,050
PM	86	202,268
PN	340	308,504
OW	62	197,011
OC	938	93,932
OY	471	92,746
Total	1,980	1,098,511

```
<?xml version="1.0" encoding="UTF-8" ?>
<sample sampleID="OW1X_00000" version="1.0" type="variableLength">
<article articleID="OW1X_00000_V001" isWholeArticle="false">
<titleBlock><title><sentence type="quasi">第２節　内外均衡の背景
</sentence></title></titleBlock>
<paragraph>
<sentence>　５３年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生
じてきたわけではない。
</sentence>
<sentence>以下では，それらの動きの重要な背景として，*snip*
</paragraph>
<cluster>
<titleBlock><title><sentence type="quasi">１．財政金融政策の効果
</sentence></title></titleBlock>
<paragraph>
<sentence>　石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向
けて運営されたが，景気の停滞が顕著となるにつれて，５０年以降５３年中に至るまで，景気
浮揚を最大の目的として運営されてきた。</sentence> ...
</paragraph>
<cluster>
<titleBlock><title><sentence type="quasi">（公共投資の拡大）
</sentence></title></titleBlock>
```

Fig. 2 Example of document structure annotation

4.2 Document structure annotation

Texts of the BCCWJ are annotated with respect to document structure. Figure 2 shows the opening part of an XML file of a sample in the OW register. Tags in this figure like <article>, <cluster>, <paragraph>, and <sentence> are all concerned with the hierarchical structure of a document, but there are also other classes of tags. Table 5 shows all tags used in the annotation of variable-length samples (see Sect. 2.3.3 above).

The type of XML document shown in Fig. 2 is called a C-XML (character-based XML) document and does not contain the results of dual-POS analysis. There is another type of XML document called M-XML (morphology-based XML) document that contains both the document structure information and the results of dual POS analysis. A crucial difference between the C- and M-XML documents is the treatment of sentence. The C-XML documents allows recursive use of the <sentence> tag within another <sentence> tag. In the example of C-XML shown in Panel A of Fig. 3, texts in the second and third lines are enclosed by the <sentence> tags, and dominated by the topmost <sentence> tag that opened

Table 5 Tags used in document structure annotation

Type of tag	Tag Name	Gloss
Sample	sample	Marks a whole sample
	sampling	Information about sampling
Hierarchical structure of document	article	Semantically coherent text written by a writer
	blockEnd	Marker of semantic boundary.
	cluster	Marks the whole text covered by a <title> tag
	titleBlock	Covers the <title> and its corresponding elements
	title	Title given to a definite area of a sample
	orphanedTitle	Title given to an indefinite area of sample
	list	Enlisted elements
	paragraph	Paragraph
	sentence	Sentence. Permits recursive application
Figures	figureBlock	Block of figure and its accompanying elements
	figure	Figure per se
	caption	Caption per se
	table	Table per se
Quotations	quotation	Element cited from outside of the current article
	citation	Element cited from other document
	source	Information about the cited document
	speech	Transcription of speech, inner speech
	speaker	Designation of speaker
	noteBody	Note and its scope
Notes	noteBodyInline	Inline note
	abstract	Abstract of article or cluster
Abstract	authorsData	Information about the author
	contents	Table of contents
Miscellaneous	profile	Profile of the authors or personae
	rejectedBlock	Shows the existence of deleted block
	verse	Marks poem, tanka, haiku, or lyrics
	verseLine	Marks one line in a verse
	ruby	Kana letters given to kanji letters to show the reading
Characters	correction	Correction of errata in the original text
	missingCharacter	Characters outside the JIS X 0213:2004 character set
	enclosedCharacter	Enclosed characters like ①,②, ©, …
	cursive	Characters in cursive style
	image	Symbols outside the JIS X 0213:2004 character set
	superscript	Superscripts
	subscript	Subscripts
	fraction	Proper fraction part of a mixed fraction
	delete	Text with strike-through line
	br	Physical line break

Table 5 continued

Type of tag	Tag Name	Gloss
	info	Information about character in the original text
	rejectedSpan	Shows the existence of deleted characters
	substitution	A character substituted by different character (s)

in the first line and closed in the fourth.[7] Note that the topmost <sentence> has its own texts. They are printed in the first and fourth lines in the panel.

Panel B of Fig. 3 shows the same document in the format of an M-XML document.[8] All texts, including the ones belonging to the topmost <sentence> in Panel A, are treated uniformly as <sentence>, and dominated by a newly introduced <superSentence> node. As shown by this example, the recursive structure of sentences in the C-XML document is transformed into a 'flat' structure consisting of only two levels, i.e., <sentence> and <superSentence>, in the M-XML document.

The 'flat' structure in the M-XML document does not reflect the syntactic structure of the original text (which is reflected in the tagging of the C-XML document), but it is preferred from a point of view of POS analysis for two reasons. Firstly, length of the text enclosed by a <sentence> tag in the M-XML document is much more moderate compared to the text enclosed by the topmost <sentence> tag in the C-XML document, which often contains several hundred SUWs, and hence is more appropriate to be the input for the POS analysis system. Secondly, it is convenient to manage the corpus data using the sentence of the M-XML documents as the unit of data management. See the user's manual of BCCWJ for more details about the two XML formats (NINJAL 2011).

4.3 Meta-information

Meta-information of the BCCWJ consists of the following four categories: bibliographical information, directory information, sampling information, and article information. Bibliographical information is the information about the original document from which the sample in question is extracted. Table 6 summarizes the contents of bibliographical information. The information recorded in fields named "Genre_1" to "Genre_3" differs according to the media of the text (namely, book, magazine, newspaper, etc.). Table 6 shows the case of books. "C-code" in Genre_3 is a book classification code used in book stores in Japan. This code classifies the supposed customer of the book (general, educated, practical, women, children, etc.), while NDC classifies the subjects of books.

Directory information deals with the data about the person or institution recorded in the Bib_author field of the bibliographical information. Table 7 shows the four

[7] The intermediate two sentences were enclosed by the <quote> tag, because they were quoted by the Japanese quotation symbols ("「" and "」") in the original text.

[8] Dual POS information is not shown here for the sake of visibility.

A：C-XML document

```
<sentence>驚きながらそう誤魔化した構治の言葉に、
<quote>「<sentence>落ちた、落ちたって言わないでよ。</sentence>
<sentence type="quasi">結構辛がってるんだから</sentence>」</quote>
言って夕美子は目を伏せ、*snip* うつむいている。</sentence>
```

B：M-XML document

```
<superSentence>
<sentence type="fragment">驚きながらそう誤魔化した構治の言葉に、</sentence>
<quote><sentence>「落ちた、落ちたって言わないでよ。</sentence>
<sentence type="quasi">結構辛がってるんだから」</sentence></quote>
<sentence type="fragment">言って夕美子は目を伏せ、*snip* うつむいている。</sentence>
</superSentence>
```

Fig. 3 Example of sentence annotation in C-XML and M-XML documents

fields of the directory information. The information of gender and birth year is obtained either from open sources or by the questionnaire that was sent to the authors as a part of the copyright processing. The meta-information will be quite valuable for the users who want to analyze the corpus from a sociolinguistic point of view.

Sampling information deals with the data about how each sample of the BCCWJ was extracted from the population. It consists of four fields: Sample_ID, Bib_ID (unique index given to the original text from which the sample was extracted), Sampling_page (page number of the original text from which the sample was extracted), and Sampling_point (information about the beginning of the sample in the page specified by the Sampling_page field).

Lastly, article information deals with the data about the unit called “article” that plays a crucial role in the specification of authorship. It happens sometimes that different parts of a single sample are written independently by different authors. Whenever it is possible to divide a sample into separate parts that are written by different authors, each part is called an ‘article.’ Article information of the BCCWJ consists of fields like Article_ID (unique ID of articles), Directory_ID (unique ID of the author of article), First_appearance (the year when the article appeared for the first time in whatever media), First_published (the year when the article is published as a book), and so forth. Note that the case of joint writing, i.e., the case when multiple authors jointly write a text which is not separable into articles, is treated as a single article.

5 Corpus release

5.1 Shonagon

Compilation work of the BCCWJ came to an end in December 2011. Currently, the corpus is publicly accessible in three different ways. The easiest way is to use a web interface program called *Shonagon*.[9] In *Shonagon*, users can search a string of up to

[9] http://www.kotonoha.gr.jp/shonagon/.

Table 6 Contents of the bibliographical information about books

Field	Gloss
Bib_ID	ID of the original text from which the sample is extracted
Title	Title of the original text
Subtitle	Subtitle of the original text
Number	Volume and number of the original text
Bib_author	Author of the original text
Publisher	Publisher of the original text
Year	Published year of the original text
ISBN	ISBN of the original text
Size	Book size of the original text
Pages	Number of pages of the original text
Genre_1	The first digit of NDC (see Sect. 2.2.1 above)
Genre_2	The first 3 digit of NDC
Genre_3	C code
Bib_author_ID	ID of the Bib_author (see above)

Table 7 Contents of the directory information

Field	Gloss
Directory_ID	ID given to the person or institution
Name	Name of the person or institution
Gender	Sex of the person
Birth year	Birth year (per decade as 1950s, 1960s etc.)

10 characters long from the whole body of BCCWJ by specifying registers and/or temporal coverage. When there are more than 500 hits, randomly selected 500 hits are shown with preceding and following contexts (each consisting of up to 40 characters). Some of the meta-information regarding bibliographical and directory information are also shown. Use of *Shonagon* is free of charge.

5.2 Chunagon

In *Shonagon*, only string search is possible. When one wants to use the results of dual POS analysis, one has to utilize *Chunagon*, another web interface program.[10] This interface is also free of charge, but the user registration by means of snail mail is required. This cumbersome process is introduced to protect the right of copyright holders.

In *Chunagon*, it is possible to search N-gram of SUW or LUW (where N is 1 ≤ N ≤ 10. Mixture of SUW and LUW is not allowed). Each of the SUW/LUW in the N-gram can be specified finely with respect to morphological information like POS (53

[10] https://chunagon.ninjal.ac.jp/.

```
キー: (品詞 LIKE "形容詞%" AND 活用形 LIKE "連用形%") AND 後方共起: 語彙素 = "た" ON 1
WORDS FROM キー AND 後方共起: 語彙素 = "です" ON 2 WORDS FROM キー AND 後方共起:
書字形出現形 = "。" ON 3 WORDS FROM キー IN core="true" WITH OPTIONS unit="1" AND
tglWords="20" AND tglKugiri="|" AND tglFixVariable="2"
```

Fig. 4 An example of a query stored in *Chunagon*

classes classified in three layers), lemma, reading of lemma, word form, written form, conjugation form (provisional, gerund, conclusive, adnominal, conditional, etc.), conjugation type and so forth. When multiple conditions of morphological information are specified, the GUI of *Chunagon* combines the conditions using logical product (i.e. AND). One way of specifying logical sum (i.e., OR) of conditions is to edit a query stored in *Chunagon*. See below in this subsection.

It is possible to limit the range of search in several independent ways. First, it is possible to specify registers from which the samples are to be searched. Second, it is possible to limit the range or search by specifying the type of sample-length; the three possible choices are fixed-length samples only, variable-length samples only, and both of them (excluding the overlap). Lastly, it is possible to make search only of the materials included in the BCCWJ-Core.

As in *Shonagon*, *Chunagon* displays only 500 hits on the screen if there are more than 500 hits. Unlike *Shonagon*, however, users can download all hit items unless the total number of hit is over 100,000.

All queries issued by the same user are automatically stored in *Chunagon* for reuse; users can access the record to reissue the same queries, or they can create new queries by editing the recorded queries. Figure 4 shows an example of a query stored in *Chunagon*. As mentioned above, users can use a logical sum (OR) while editing the queries.

As can be seen from the figure, the syntax of the query language resembles that of SQL. This query defines a search for the cases where an adjective in their adverbial form is immediately followed by an auxiliary verb た "ta" ('PAST'), which is in turn followed immediately by another auxiliary verb です "desu" (polite form of copula) before a sentence boundary marked by a period. It is also specified that the range of the search is limited to the BCCWJ-Core, and the SUW is specified as the morphological unit. Similar queries were used in the analysis of i-adjectives reported in Sect. 6.5 below.

Lastly, users can issue multiple queries at a time when he/she is dealing directly with the query language (rather than making a query using the GUI mentioned above). This contributes greatly to the speeding-up and the reproducibility of corpus analysis.

5.3 DVD-release

It is impossible to make access to the full contents of the BCCWJ even in the environment of *Chunagon*. Users who want to utilize the whole information in the BCCWJ described in Sects. 2 and 4 should utilize DVD-release version of the corpus. In this version, all C-XML/M-XML documents (see subsection 4.2 above), meta-information data, and the data table used in *Chunagon* (in the TSV format) are packaged in two DVDs. This is literally the whole body of the BCCWJ.

Note that the DVD-release version does not include any environment for corpus search. Users should construct the environment of corpus query by themselves. For most users, use of RDBM like MySQL is recommended. Note also that with regard to the DVD-release of the BCCWJ, the price of permanent academic license is 50,000 JPY.[11]

6 Some preliminary analyses

In the rest of this paper, results of some preliminary analyses of the BCCWJ are presented. These analyses are concerned with the BCCWJ-Core, and, the focus of analyses is placed on the diversity of texts in the corpus as observed across registers.

6.1 POS distribution

Many preceding studies report that the POS distribution differs systematically depending on the type of texts (see Kabashima and Satake 1978 among others). Figure 5 compares the distributions of four main content-word classes (noun, verb, adverb, and adjective) in SUW. Note that two types of adjectives in Japanese, namely i-adjectives (i.e., adjectives ending in /i/), and na-adjectives (adjectives ending in /na/), are pooled into one class.

Registers like PN (newspaper) and OW (white paper) whose main function is the transmission of facts are marked by a higher ratio of nouns and lower ratio of adverbs, while registers like OC (bulletin-board) and OY (blog) whose main function consists in the expression of subjective opinion are marked by a lower ratio of nouns and higher ratio of adverbs.

6.2 Distribution of word-classes

UniDic has a field for word-class information, i.e., Japanese (native Japanese), Sino-Japanese (historical loans from Chinese), loans (modern loans from English and other European languages), and their hybrids. Punctuation marks and various alphabetical abbreviations like “OS” and “HDD” are classified as belonging to the fifth word-class: symbol.

Figure 6 shows the distribution of these word-classes in the SUW. The most striking genre-related difference seems to be the ratio of Sino-Japanese vocabulary. Registers OW and PN are the highest, while OC and PB are the lowest. Also, it is interesting to see that the sum of the ratios of Japanese and Sino-Japanese remains nearly constant (88–92 %) across all registers. These tendencies coincide fairly well with the observations reported in Koiso et al. (2009) who examined the distribution of word-classes in the BCCWJ during its construction.

[11] Application information can be found at http://www.ninjal.ac.jp/corpus_center/bccwj/subscription.html.

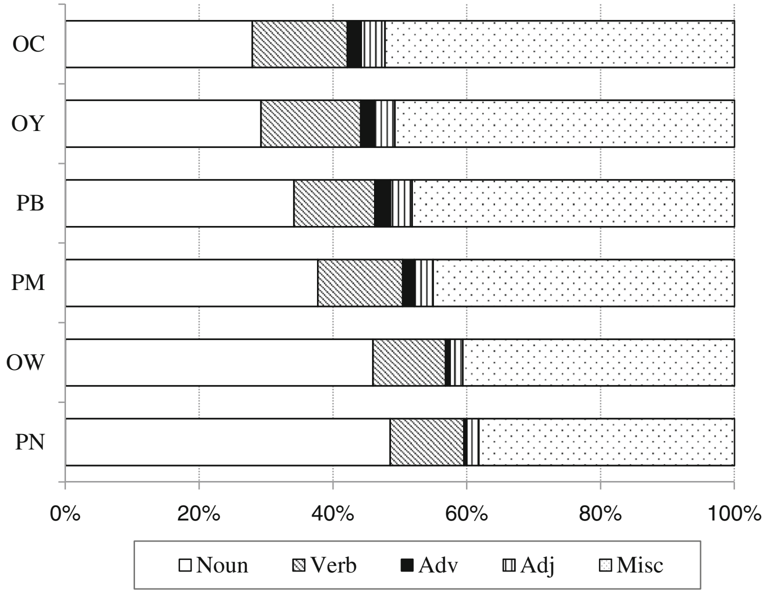

Fig. 5 Difference of POS distribution due to registers

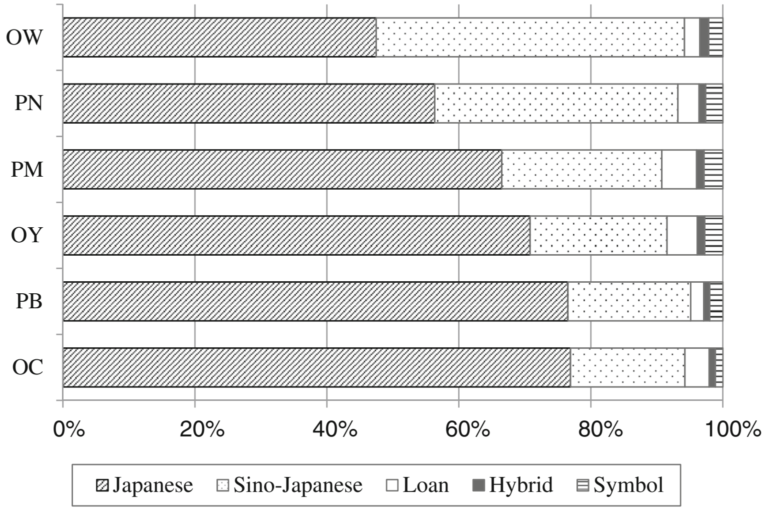

Fig. 6 Distribution of word-classes

6.3 Entropy of orthographic variations

In Sect. 1, it is suggested that newspaper articles are a class of text where orthographic variations are extensively suppressed; the implication is that there are other registers where the variations are relatively larger. This hypothesis is tested

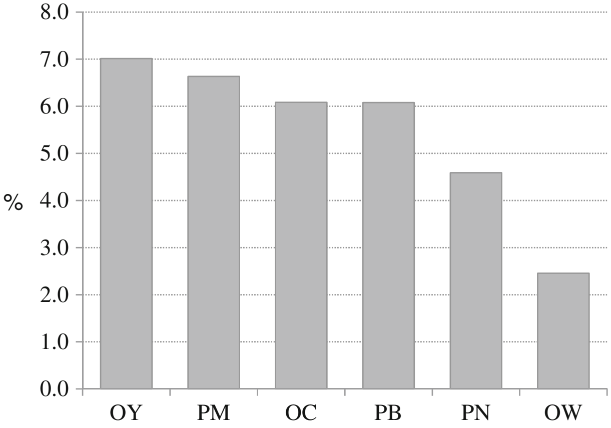

Fig. 7 Ratio of variable nouns

here, by quantifying the degree of orthographic variation of the nouns in the six registers of the BCCWJ-Core.

To achieve this objective, two analyses are conducted. First, the ratio of the total number of variable nouns (i.e., the nouns that have more than two ways of being written) to the total number of all nouns (number in terms of types rather than tokens) was computed for each register. Figure 7 shows the results. The register that showed the lowest ratio is OW (white papers), and the PN (newspapers) is the second lowest. On the other hand, registers OY (blog) and PN (magazines) showed the highest ratio of variable nouns.

The second analysis is the computation of information entropy. Entropy in information theory is a mathematical measure of the uncertainty associated with a random variable. The entropy of coin tossing is equal to 1.0 bit, while the entropy of one cast of dice is about 2.59 bit. Larger entropy implies larger uncertainty of the random variable (in this case, the orthography of a noun). Figure 8 is a composite bar graph showing two results of entropy computation simultaneously.

The wide filled bars that locate behind the narrow dotted bars show the mean entropy of all variable nouns in each register, and the axis of entropy is on the left side of the figure. The narrow dotted bars show the mean entropy of all nouns (covering both variable and non-variable ones) in each register, and the axis is on the right side. In both cases, OW is the lowest and the PN is the second lowest. On the other hand, the highest register differs depending on the ways of computation. OC and OY are the highest with respect to variable nouns, while OY and PM are the highest with respect to all nouns. The fact that OY is always among the highest coincides well with the result shown in Fig. 7. And, as predicted, PN belongs to the class of registers where orthographic variation is very low (if not the lowest).

6.4 Sentence length

So far, text diversity was examined mostly with respect to morphological properties. Two syntactic properties are examined below. Probably, the simplest measure of

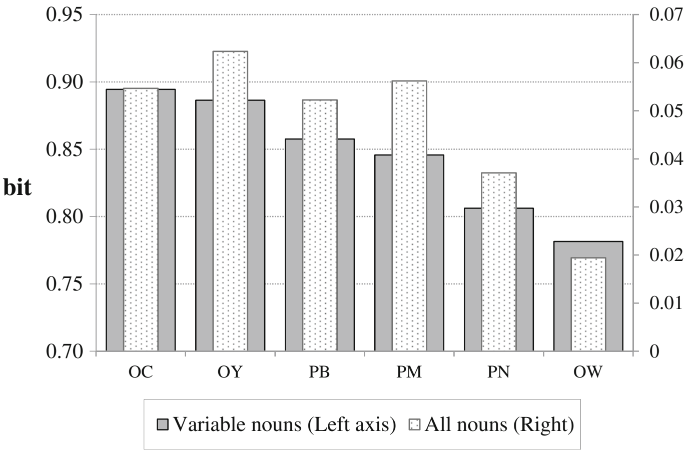

Fig. 8 Mean entropy of nouns

sentential text diversity is the comparison of sentence length. Figure 9 is the comparison of mean sentence length as measured by the number of SUW in a sentence across registers. Punctuation marks are removed from the data. Sentences in the OW register are by far the longest. On the other hand, sentences are the shortest in the internet-related registers (OC and OY). The result of a one-way ANOVA is significant (DF = 5, F = 290.32, $P < 0.001$). Pairwise t tests with Bonferroni correction reveal that differences of all combinations of two registers are significant at the 0.001 level, except for the pairs of PB and PN, and, PM and PN.

6.5 Adjective predicate

The second syntactic property to be examined is the structure of i-adjective predicates. In Japanese, four POS categories can constitute a predicate, namely, nouns, na-adjectives, i-adjectives, and verbs. In the standard description of the Japanese grammar (see Teramura 1982 among others), it is acknowledged that nouns and na-adjectives need to be followed by a copula (“da” or “desu”) to constitute a predicate, while i-adjectives and verbs constitute predicates by themselves without the presence of a copula. In everyday use of the language, however, i-adjectives in the predicate position are often followed by a copula “desu”. Accordingly, “ano hon wa omoshiroi” (‘that book is interesting’, “ano” ‘that’, “hon” ‘book’, “wa” TOPIC, “omoshiroi” ‘interesting’) and “ano hon wa omoshiroi desu” are both observed in the real usage.

Figure 10 compares the relative frequencies of two types of i-adjective predicates. The legend “Adj” stands for the cases where i-adjectives constitute a predicate by themselves, and, “Adj + desu” stands for the predicate with a copula.

Here, three cautions are to be presented. First, cases where i-adjective + “desu” is followed by phrase-final particles like “ne”, “yo”, “ka” are all excluded. Second, cases where i-adjectives are followed by the conjectural form of copula “desho” are

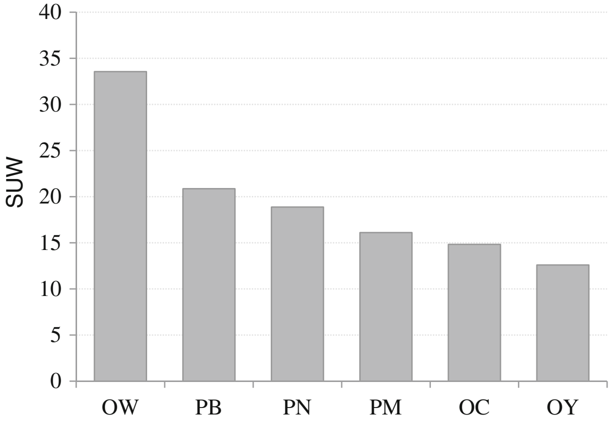

Fig. 9 Comparison of mean sentence length

all excluded. These two cases have to be removed from the data because these predicates are widely recognized as grammatical in the traditional description of Japanese syntax.

Third, the present data is based upon the analysis of the cases where i-adjectives constitute a predicate at sentence-final position. Subordinate sentences and coordinate sentences in non-final positions ending in i-adjective predicates are not included in the data.

The figure shows clearly that use of the “Adj + desu” predicate is heavily concentrated in two internet-related registers, OC and OY. Especially in OC, the predicate with a copula is in the majority. In registers PN, PB, and OW, on the other hand, no instance of “Adj + desu” predicate is found as far as the BCCWJ-Core is concerned. Analysis of the whole BCCWJ shows the same overall pattern, but a handful of “Adj + desu” predicates are found in registers PB, PN, LB, OB, and OT. The ratio of the “Adj + desu” in these registers is generally less than 1 % (Maekawa 2012).

6.6 Automatic classification of registers

Lastly, automatic classification of text registers is conducted using the measures of text diversity presented in the previous subsection. A subset of the BCCWJ-Core is constructed such that six registers have 50 samples each. Then, each sample in the subset is characterized by a vector comprised of the following eight variables: sentence length, the ratios of nouns, verbs, adverbs, and adjectives in the POS distribution, the ratios of Japanese words, Sino-Japanese words, and loan words in the word-class distribution. All variables are z-normalized before the analysis. POS ratios of adverbs and nouns, and word-class ratio of loan words are log-transformed before the normalization.

Support vector machines, or SVM, were used as the classification tool. Function svm of the e1071 library of the R language (ver. 2.15.1) was used for analysis with

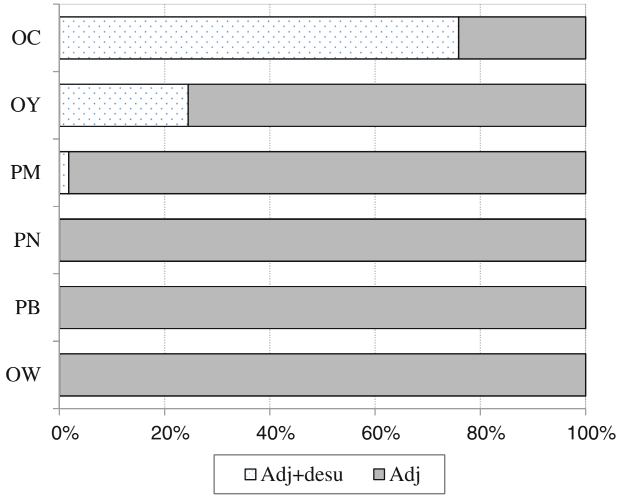

Fig. 10 Relative frequencies of two types of i-adjective predicates

Table 8 Result of register classification by SVM

	OC	OW	OY	PB	PM	PN
OC	42	0	5	2	1	0
OW	0	49	0	0	0	1
OY	3	0	40	3	0	4
PB	2	0	0	47	1	0
PM	0	1	3	12	27	7
PN	0	0	0	3	5	42

the default setting of kernel (radial) and type (C-classification). SVM parameters were searched within the ranges between 10^{-5} and 10^{-2}, and, 10^{-2} and 10^{2} respectively for gamma and cost.

Table 8 shows the cross-tabulated results of the tenfold cross-validation. The rows and columns correspond respectively to correct and classified registers. The overall average rate of success is 0.83, and is much higher than the baseline (1/6–0.17). It suggests strongly the usefulness of some of the measures of text diversity presented in this section as the cue for the characterization of texts in the BCCWJ. At the same time, it turns out that the register PM is the most difficult to classify. A straightforward interpretation of this fact is that magazines belong to a heterogeneous register covering a wide range of texts overlapping considerably with the typical texts of other registers, especially PB and PN. Another tendency worth mentioning is the proximity between OC and OY, but this is not as salient a tendency as that of PM.

7 Conclusion

BCCWJ is designed to be a reliable reference corpus of present-day written Japanese with as much representativeness as possible. Compilation of the BCCWJ terminated successfully in December 2011 with almost no delay in the original schedule. Pilot evaluations of the corpus revealed that the texts in the corpus are much more diverse than the texts previously analyzed in most of the corpus-linguistic studies of Japanese, i.e., the newspaper articles.

As stated above, the BCCWJ is publicly accessible in three different ways. As of October 2013, *Shonagon* has more than 297,000 accumulated visitors. *Chunagon* has more than 1,500 licensees. And there are more than 250 licensees of the DVD-release version. It is the belief of the present authors that the new era of corpus-based analysis of present-day Japanese is now being opened by the users of the BCCWJ.

Acknowledgments The authors are grateful for the more-than-15-thousands copyright holders who gave us permissions to use their materials in the BCCWJ. They are also grateful for their former colleagues whose names are not included in the authors' list. They played important roles in the compilation of the BCCWJ. Compilation of the BCCWJ was supported by the grant-in-aid for scientific research for priority-area research (Grant No 18061009, 2006–2010) to the first author and the research budget of the former National Institute for Japanese Language. At the same time, this paper is the outcome of the collaborative research project "Foundation of Corpus Japanese Linguistics" of the current National Institute for Japanese Language and Linguistics. Lastly, the authors are grateful for John Whitman and Donna Erickson who kindly checked various versions of this paper and provided us with many valuable comments.

References

Asahara, M., & Matsumoto, Y. (2000). Extended models and tools for high-performance part-of-speech tagger. In *Proceedings of the 18th international conference on computational linguistics (COLING 2000)*, pp. 21–27.

Den, Y., Nakamura, J., Ogiso, T., & Ogura, H. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceeding of the 6th Language Resources and Evaluation Conference (LREC 2008)*, pp. 1019–1024.

Fujiike, Y., Ogura, H., Konishi, H., Ogiso, T., Koiso, H., Uchimoto, K., & Kozawa, S. (2010). Gendai nihongo kakikotoba kinko kopasu niokeru chotan'ikaiseki no shinchokujokyo. In *Tokutei ryoiki kenkyu nihongo kopasu heisei 21 nendo kokai wakushoppu yokoshu*, National Institute for Japanese Language and Linguistics, pp. 93–99.

Kabashima, T., & Satake, H. (1978). *Shin bunsho kogaku: Hyougen no kagaku*. Tokyo: Sanseido.

Koiso, H., Ogiso, T., Ogura, H., & Miyauchi, S. (2009). Kopasu ni motozuku tayo na janru no buntai hikaku: Tantan'i joho ni chumoku shite. In *Proceedings 15th annual meeting of the Association for Natural Language Processing*, pp. 593–597.

Kozawa, S., Uchimoto, K., & Den, Y. (2011). BCCWJ ni motozuku chu-chotan'i kaiseki tsuru. In *Tokutei ryoiki kenkyu nihongo kopasu heisei 22 nendo kokai wakushoppu yokoshu*, National Institute for Japanese Language and Linguistics, pp. 331–338.

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of EMNLP, 2004*, pp. 230–237.

Kurohashi, S., & Nagao, M. (1998). Building a Japanese parsed corpus while improving the parsing system. In *Proceedings 1st international conference on language resources and evaluation*, pp. 719–724.

Kurohashi, S., Nakamura, T., Matsumoto, Y., & Nagao, M. (1994). Improvements of Japanese morphological analyzer JUMAN. In *Proceedings international workshop on sharable natural language resources*, pp. 22–28.

Maekawa, K. (2003). Corpus of spontaneous Japanese: Its design and evaluation. In *Proceedings ISCA and IEEE workshop on spontaneous speech processing and recognition, (SSPR 2003)*, pp. 7–12.

Maekawa, K. (2012). Keiyoushi + desu jutsugo no seiki yoin ni tsuite no junbi teki kosatsu. In *Proceedings 1st Japanese corpus linguistics workshop*, pp. 211–220.

Maekawa, K., Koiso, H., Furui, S., & Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proceedings 2nd international conference on language resources and evaluation*, pp. 947–952.

Matsuda, K. (Ed.). (2008). *Kokkaikaigiroku o tsukat ta nihongo kenkyu*. Tokyo: Hituji.

NINJAL (Ed.) (2011). *BCCWJ riyo no tebiki* (Electronic document delivered with the DVD-release version of the BCCWJ).

Teramura, H. (1982). *Nihongo no shintakkusu to imi* (Vol. 1). Tokyo: Kuroshio.

Uchimoto, K., & Isahara, H. (2007). Morphological annotation of a large spontaneous speech corpus in Japanese. In *Proceedings 20th international joint conference of artificial intelligence*, pp. 1731–1737.