KOTONOHA and BCCWJ: Development of a Balanced Corpus of Contemporary Written Japanese

Kikuo Maekawa
Department of Language Research
The National Institute for Japanese Language
kikuo@kokken.go.jp

Abstract

The National Institute for Japanese Language (NIJL) has launched a long-term language corpus development initiative aiming at the development of a super-corpus called KOTONOHA, which is consisting of a multitude of independent corpora. Among the constituent corpora of KOTONOHA, the one that bears the most urgent need is a large-scale balanced corpus of the present-day written Japanese. Construction of such a corpus is currently underway under the auspice of NIJL and financial support of grant-in-aid for scientific research from the MEXT. This paper describes the basic design issues of the balanced corpus called BCCWJ. The BCCWJ consist of three component sub-corpora differing in the nature of statistical populations, i.e, 'production', 'circulation', and 'non-population' sub-corpora. The first two sub-corpora represent the production and reception aspects of published written Japanese, while the last sub-corpus is aggregate of various mini corpora developed for specialized research and language planning purposes.

1. Introduction

In Japanese linguistics, as in many other languages, corpus-based analysis has been becoming more and more prevailing. However, it is widely recognized by specialists that the lack of balanced, or reference, corpus of the contemporary Japanese is the fundamental problem of the Japanese corpus linguistics. Most of the corpus-based studies of contemporary Japanese are based upon the analyses of text database of newspaper articles or crawling of the Internet text.

Lack of balanced corpus imposes two mutually related problems on the study. Most of newspaper articles are written by newspaper writers who are very much aware of the established writing style (and orthography) of newspaper article. Accordingly, it is the genre of text where variations of all sorts are suppressed to the minimum level.

On the other hand, the results of internet crawling using search engines like Google and Yahoo are very much likely to include texts encompassing wide ranges of text. It is also expected that considerable amount of linguistic variations are to be observed. It is, however, very difficult, if not impossible, to conduct analyses of style difference and/or linguistic variations using the results of internet crawling, because the results are not classified in terms of text genres, for one thing, and the amount of the text can often be excessively large to be classified by hand. There is also problem of the reproducibility. It is widely recognized by those who are working especially with Google engine that the output to the same set of key words differs considerably from time to time, even from one instance to another.

To solve these and other difficulties in Japanese corpus linguistics, National Institute for Japanese Language (NIJL, hereafter) has recently launched a new corpus compilation project in the spring of 2006, aiming at public release of Japan's first balanced corpus in the year of 2011. This balanced corpus is deemed to be a component of super-corpus called KOTONOHA that covers the full range of modern- and contemporary Japanese. In the rest of this paper, I will first describe NIJL's KOTONOHA initiative and its component corpus CSJ. Then I will go on to present the design issues of the new balanced corpus.

2. Background

2.1. KOTONOHA

NIJL has a long-term corpus development initiative for the modern and contemporary Japanese, which is known as the KOTONOHA project. KOTONOHA is a cover-term for a multitude of corpora covering the whole range of modern Japanese. Figure 1 shows the current status of KOTONOHA. The abscissa of the figure is the time axis that covers schematically the time from the beginning of the Meiji era (when Modern Japan started) up to the present. The ordinate of the figure represents schematically the different genres in the language. The upper and lower halves of the ordinate stand respectively for written and spoken language, and each of them are divided into different subgenres; books, newspapers, and magazines in the case of written language, and, monologue, dialogue, and chat in the case of spoken language.

Two ellipses in the figure stand for the two component corpora of KOTONOHA that have already been publicly available. On the one hand, there is *Taiyo Corpus* that covers the texts of Taiyo magazine published in the years 1895-1925, which was a general-interest magazine read by wide range of readers of that time. There is also *Corpus of*

Spontaneous Japanese (CSJ) that records mostly spontaneous monologue of the contemporary Japanese (recorded in 1999-2002).

Rectangles in the figure stand for KOTONOHA's component corpora that are to be compiled in the coming years, and, the one located at the upper right corner of the figure is the most important for the present paper. This corpus stands for a balanced corpus of contemporary written Japanese containing at least 100 million words, and is named *Balanced Corpus of Contemporary Written Japanese*, or BCCWJ.

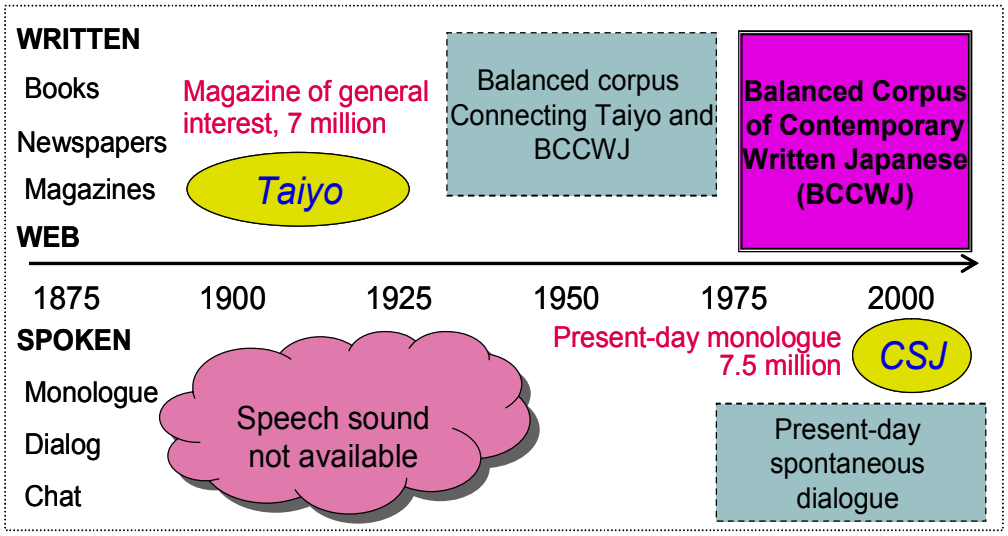

Figure 1. Schematic representation of KOTONOHA. Ellipse and box stand respectively for corpora publicly available and to be compiled. BCCWJ locates at the upper right corner. Dotted boxes are those to be compiled after the completion of the BCCWJ.

2.2. Pilot corpus and application for national grant

In the summer of 2004, shortly after the public release of the CSJ, and shortly before the completion of the *Taiyo Corpus*, a volunteer group of NIJL researchers, who were mostly involved in one of the two corpora mentioned above, started conjointly a volunteer project that aimed at the evaluation of the possibility of compiling a balanced corpus of the present-day Japanese. They spent a whole year of 2005 for the construction of a pilot balanced corpus containing about 1.2 million words for the evaluation of various design issues. At the same time, they formed a nation-wide research team for the compilation and analysis of BCCWJ, and applied for a MEXT (Ministry of Education) grant-in-aid for priority area scientific research.

In the spring of 2006, NIJL established a new research team organized specifically for BCCWJ, and the team started compiling the corpus aiming at the public release of

the corpus in the year of 2011. Soon after its take off, the BCCWJ project was blessed by the MEXT's announcement of the acceptance of the priority area program as a five-year (2006-2010) national project, known as the 'Japanese Corpus' project (See the following URL for more details: http://www.tokuteicorpus.jp). In the rest of this paper, I will first present briefly the structure of the CSJ and then go on to discuss design issues of the BCCWJ. These are the two main component corpora of the present-day Japanese in KOTONOHA.

3. The CSJ

CSJ, or *Corpus of Spontaneous Japanese*, is a richly annotated large-scale corpus of spontaneous monologue designed primarily for the study of automatic speech recognition (ASR). More than 660 hours of speech signal contained in the corpus are all transcribed and POS annotated. See Table 1 below.

Use of CSJ in the ASR of spontaneous Japanese increased drastically the correct recognition rate. Word recognition rate jumped up from 45% to 80% as the combined effect of new language- and acoustic-models that were automatically learned from CSJ and improvements in the pattern-recognition algorithms for ASR like speaker adaptation.

ASR is not the only application domain of the CSJ, however. The corpus is also designed for the study of language variation. For this particular purpose, CSJ has a special subset, called the CSJ-Core, to which the cost of corpus annotation is concentrated (See Tables 2-3 below).

Table 1. Size of the CSJ

N of running words *	7,525,125
N of different speakers	1,417
N of talks	3,302
Total hour of speech	662

* Counted in terms of SUW (Short Unit Word)

As can be seen in Table 2, more than 95% of the talks were devoted for monologue, and the resulting 5% were devoted for dialogue and read speech. It is important to note that systematic difference of speaking style is observed not only between monologues and dialogues, but also between the two main monologue types, i.e, Academic Presentation Speech (APS) and Simulated Public Speaking (SPS). APS is the live

recording of academic presentation talks done in various academic meetings covering humanities, engineering, and social sciences. SPS comprises public speech done by recruited layman speakers on various everyday topics (like 'the town I live in' or 'the most joyful memory of my life'). There is also considerable difference of speaking style depending on the talkers.

Table 2. Type of talks in the CSJ.

TYPE OF TALKS	MODE	N FILE	N SPKER	HOUR
Academic Presentation Speech (APS)	Monologue	987	819	274.4
Simulated Public Speaking (SPS)	Monologue	1,715	594	329.9
Public Lectures (PL)	Monologue	19	16	24.1
Interview on APS	Dialogue	10	(10)	2.1
Interview on SPS	Dialogue	16	(16)	3.4
Task-oriented dialogue	Dialogue	16	(16)	3.1
Free dialogue	Dialogue	16	(16)	3.6
Reread speech	Monologue	16	(16)	5.5
Read speech	Monologue	507	(248)	15.5

Table3. Annotation of CSJ

ANNOTATION	Whole CSJ	CSJ-Core
Speech signal	✓	✓
Speaker info	✓	✓
Two-way Transcription	✓	✓
Two-way POS analysis	✓	✓
Clause boundary Info	✓	✓
Impression rating score	✓	✓
Segmental label	N.A.	✓
Intonation label	N.A.	✓
Dependency analysis	N.A.	✓
Topic boundary info*	N.A.	✓

* Available only for a subset of the CSJ-Core. N.A. stands for 'Not Available.' POS analysis and clause boundary information in the Core are high in accuracy because they are manually checked. Impression rating score for the Core are rated by multiple raters.

Transcription system of the CSJ is called two-way transcription because it provides us with orthographic and phonetic transcriptions. The orthographic transcription is given in Kanji (Chinese logograph) and Kana (Japanese syllabary) and is used for information retrieval purposes, while the phonetic transcription is given only in Kana characters and is used for the study of phonetic variations.

The POS (part-of-speech) information of the CSJ is also provided in two different ways; in terms of Short Unit Word (or SUW) and Long Unit Word (LUW). SUW is relatively short unit that approximates, by and large, the size of dictionary items that were used traditionally in Japanese dictionaries (*Kokugo Jiten*). On the other hand, LUW is the unit suitable to represent compound words in Japanese. For example, the Japanese name of the NIJL /kokuricukokugokenkyuHzyo/ consists of four SUWs ({kokuricu} 'national', {kokugo} 'national language', {kenkyuH} 'research', and, {zyo} 'institution'), but it makes only one LUW.

The 'impression rating score' annotation was introduced to evaluate differences in speaking style. All monologue talks in the CSJ were evaluated by human raters at the time of recording with respect to various subjective dimensions, including speaking rate, speaking style, spontaneity of talks, use of prepared text, etc. Speaking style, for example, is evaluated using the scale of 1 to 5, higher the number more formal the speech.

After its public release in the spring of 2004, more than 330 copies of the CSJ has delivered for various research institutions including many private companies, both home and abroad. Furui et al (2005) summarizes the effectiveness of the CSJ in the field of speech engineering, and Maekawa (2005, 2006) present examples of the study of language variation using the CSJ. See the following URL for more results of the preliminary analyses of the CSJ: http://www2.kokken.go.jp/csj/public/index.html

4. Design of the BCCWJ

One of the most important, and controversial, issues of the design of balanced corpus is that of securing representativeness (See Biber 1993 and Kennedy 1998 among others). Needless to say, there is no simple solution to this issue; there are multiple ways of securing corpus representativeness, depending on the purpose of corpora.

In the case of BCCWJ, the NIJL research group spent almost the whole year of 2005 discussing this issue before arriving at the conclusion that we had better adopt two different kinds of representativeness and construct three different sub-corpora. Figure 2 represents schematically the inner structure of BCCWJ. The corpus consists of three different sub-corpora, viz., 'production', 'circulation', and 'non-population' sub-corpora. The following subsections will be devoted for the introduction to each of these sub-

corpora. The issues of specimen length and corpus size will also be discussed.

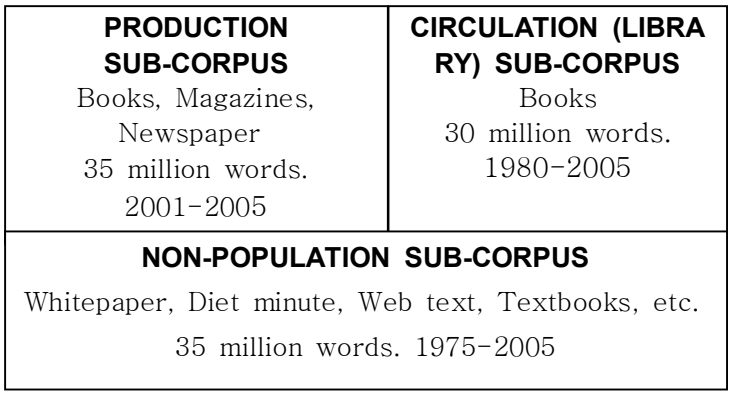

Figure 2. Three component sub-corpora of the BCCWJ.

4.1. Production sub-corpus

The upper left-hand sub-corpus of Figure 2 is called 'production' sub-corpus. As suggested by its name, this sub-corpus represents the production, as opposed to the reception aspect of contemporary written Japanese. The sub-corpus consists of samples extracted randomly from the statistical population covering the whole body of books, magazines, and newspapers published in the years 2001-2005.

It is important to note that the population for statistical sampling is defined explicitly using the data sources that are publicly available; *J-BISC* (Japan Biblio Disc) for books, and *Periodicals in Print in Japan* for magazines, for example. To be more exact, we estimated the total number of characters involved in the population, and drew samples from the population in the way that each character in the population had the same chance of being sampled. We estimated the number of characters rather than words, because word boundary in Japanese is heavily theory dependent and hence is not reflected in ordinary orthography.

At this point, note especially that the composition ratios among genres (i.e. the ratio among samples of books, magazines, and newspapers) were determined on the basis of publicly available data mentioned above. This makes crucial difference from the design of corpora like the *Brown Corpus* and *British National Corpus*, where the composition ratios of various genres were determined subjectively by specialists of English without making reference to objective data (See http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM and http://www.natcorp.ox.ac.uk/ respectively).

The total size of the sub-corpus is supposed to be about 34.7 million words (see section 4.5 for details), and according to the latest estimation, 74.1, 16.1, and 9.8% of

the production sub-corpus are to be devoted to samples of books, magazines, and newspapers respectively.

Lastly, the production sub-corpus is designed to cover the most recent five years in the past so that it will represent the most recent status of the contemporary written Japanese. See the discussion at the end of the following subsection.

4.2. Circulation sub-corpus

The production sub-corpus will provide us with statistically exact information about how written Japanese are published. At the same time, however, it is possible to conceive of balanced corpora that represent aspects of language usage other than the production aspect. There might be users who find it more important to obtain information about the reception of a language than the production.

From the point of view of such users, it is desirable to have a corpus in which a book that sold one million copies is weighted ten thousand times of a book that sold only one hundred copies. It is, however, very difficult to obtain reliable information about the reception aspect of language. It often turns out to be the case that publicly available data about the sales of books and magazines are absent or incorrect.

Our second sub-corpus, which is located at the upper right-hand of Figure 2, is called 'circulation' or 'library' sub-corpus. This sub-corpus contains samples taken from those books that could be regarded to have been received by certain amount of readers. To achieve this objective, we defined the statistical population of this sub-corpus as the books registered in multiple public libraries in the Tokyo Metropolis. Due to the courtesy of Tokyo Metropolitan Library, we could analyze the data about the books registered in all the public libraries of the Tokyo Metropolis. It turned out that more than 10 million books are registered in the libraries, and the books consisted of 1,064,186 different books.

Accordingly, if we define the population of the sub-corpus as the books registered at least in one library, the population will consists of about one million books. Alternatively, if we define the population as the books registered at least in five libraries, the population will have 660,516 different books, and if we define the population as the books registered at least in ten libraries, the population will have 483,569 different books. The population defined in this way will be consisting of those books that were 'received' at least by certain number of readers. The size of the circulation sub-corpus will be about 30 million words (see 4.5 below).

There is also a difference between the production and circulation sub-corpora with respect to the coverage of time. In the case of circulation sub-corpus, in principle, only the books to which ISBN (International Standard Book Number) is assigned are to be

involved in the population. Roughly speaking, the population thus defined will cover the last one quarter century, i.e. the years after 1980 when ISBN started to be adopted by Japanese publishers. Although the sub-corpus is not designed as a diachronic corpus, it will provide information about the recent changes in the Japanese language. It is also important to note that the period of time covered by the sub-corpus is also the period of time when Japanese started to be influenced by the computer processing of Kanji and Kana characters, whose beginning can be traced back to the establishment of the old version of the JIS Kanji code in 1978.

4.3. Non-population sub-corpus

The third sub-corpus, which is located in the bottom of Figure 2, is called 'non-population' sub-corpus, because this sub-corpus is the aggregate of various special-purpose mini corpora that are not necessarily sampled using well-defined statistical population.

Currently, the sub-corpus contains mini corpora of the governmental white paper containing about five million words, the Internet text (Yahoo! Japan's bulletin board *Chiebukuro*) containing about five million words, and, the Minutes of Japanese National Diet (covering both the Upper House, or *Sangi'in*, and Lower House, or *Shuugi'in*) containing about five million words. Note that all these mini-corpora are sampled corpora rather than full text corpora. The whole body of *Chiebukuro* data, for example, is estimated to contain more than 180 million characters that correspond to about 100 million words.

In addition to the mini corpora mentioned above, we plan to construct mini corpora of Japanese textbooks. The textbook mini-corpora will include text sampled from the whole textbooks used in elementary-, junior-high-, and high-schools.

It is important to note that these mini corpora are linked to particular research activities in the NIJL and the priority-area 'Japanese Corpus' project. For example, the textbook mini corpus will be utilized to make the frequency list of the words used in Japanese schools. The comparison between the word frequency lists of existing textbooks and that of the whole BCCWJ will provide us with precious information for the design of basic vocabulary for educational purposes.

4.4. Size of specimen

In corpus linguistics, there is often a discussion about the pros and cons of a sampled corpus (as opposed to a whole-text corpus). For example, Kennedy (1998) points out that sampled texts may not be suitable for some kinds of studies like stylistics and discourse analysis. In the case of the BCCWJ, however, it is virtually impossible to

construct a whole-text corpus, because nearly all materials of the corpus are protected by the copyright law. So the real remaining problem is that of the length of specimen.

Based upon our prior experiences on the compilation of word frequency list using random sampling, and computer simulation experiment about the relationship between the specimen length and the coverage of different words in the population, we arrived at the conclusion to utilize two sorts of specimen differing in length. One of them is called 'fixed-length specimen' consisting of 1000 characters. The other is called 'variable-length specimen' and covers the minimum discourse structural unit, which corresponds most usually to units like section or chapter. According to the analysis of the pilot corpus, average lengths of variable-length specimen were 3900, 3000, and 1000 characters long for books, magazines, and newspapers articles respectively.

Also, we make it a rule that the size of a specimen does not exceed 10000 characters. This restriction is necessary, because there can be samples that have no clear structuring (for example very long philosophical text without any segmentation into sections or chapters).

In samples taken from books, fixed-length specimens are usually included in the corresponding variable-length specimens. But in the case of newspapers articles, it is not usually the case. Variable-length specimens are often shorter than the corresponding fixed-length specimens, because the lengths of newspaper articles are often shorter than 1000 characters. When it is the case, the fixed-length specimen will include the beginning part of the article that immediately follows the article in question.

Figure 3 shows how we determine two types of specimens in the case of newspaper whose article length exceeds 1000 characters. The circle indicates so-called 'sampling point', i.e. the letter chosen randomly

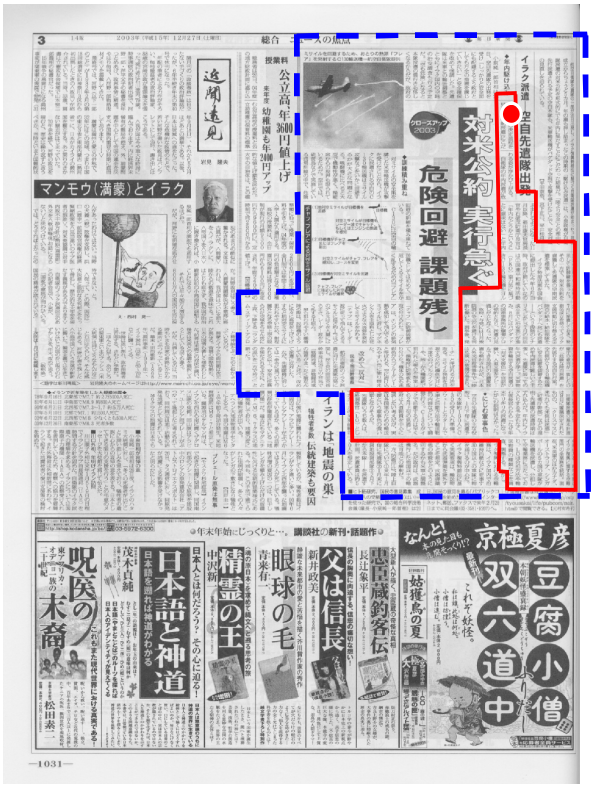

Figure 3. Example of the fixed- and variable-length specimen in the case of a newspaper article. Letters used in figures and tables are not included in the specimen. Advertisements are also neglected.

from the population. Fist we take fixed-length specimen, by choosing 1000 characters starting from the letter that locates at the beginning of the sentence that includes sampling point. Then we take the whole newspaper article as the variable-length specimen. In Figure 3, areas enclosed by the real and dotted lines represent fixed- and variable-length specimen respectively.

Variable-length specimen will be appreciated by corpus users who are interested in the study of stylistics and/or discourse structure. On the other hand, users who want to obtain statistical information as exact as possible about the population will appreciate fixed-length specimen for statistical inferences. Statistically exact inference is indispensable when, for example, we compile a frequency list of Kanji characters, which will provides the very basis of the compilation of the national Kanji lists for educational and ordinary usage purposes ('*Kyooiku*' and '*Jooyoo*' Kanji lists). The examination of the latter ('*Jooyoo'*) Kanji list is regarded to be one of the most urgent issues in the language planning of the present-day Japanese language.

4.5. Size of sub-corpora

As mentioned above, the sizes of production and circulation sub-corpora are supposed to be 34.7 and 30.0 million respectively. These sizes were determined as follows. First of all, we determined the total number of words involved in the fixed-length samples to be 10 million. According to our prior experiences of lexicographic statistical surveys, this is the minimum requirement for statistical inference about the population.

Once this is fixed, the number of fixed-length specimen can be estimated in the following way. First, we estimated the average number of word in a fixed-length specimen to be about 588 by calculating 1000/1.7, where 1.7 is the average length (in terms of the number of characters) of a SUW (see section 3 above) estimated by use of the pilot corpus. Second, the number of fixed-length specimen required to reach the total of 10 million words is supposed to be about 17000 by calculating 10000000/588. And third, the total number of words in the book part of the production sub-corpus is estimated to be about 29 million words by computing 17000*0.741*3900/1.7, where 0.741 is the ratio of book samples in the sub-corpus, 3900 is the averaged character length of variable-length specimen of books, and 1.7 is the average length of a SUW. In this way, we can estimate the size of magazine- and newspaper-parts of the sub-corpus to be 4.8 and 0.98 million respectively. See Table 4 for the result of estimation. See also section 3.4 for averaged length of variable-length specimen.

As for the circulation sub-corpus, we have not fixed its size yet, but it is highly probable that we construct a sub-corpus whose size is the same as that of the book-part of the production sub-corpus, i.e., about 30 million words. In this case, the total size of

the production and circulation sub-corpora will be about 65 million words. Lastly, the size of the non-population sub-corpus has not been fixed either. It has to contain 35 million words at the very least to fulfill our official vow about the size of the BCCWJ.

Table 4. Size of production sub-corpus

GENRE	RATIO[%]	N SAMPLE	WORD SIZE (SUW)
Book	74.1	12,604	28,915,000
Magazine	16.1	2,730	4,818,000
Newspaper	9.8	1,666	980,000
Total	100.0	17,000	34,713,000

4.6. Distribution of samples

Tables 5 and 6 show how the samples in production sub-corpus are distributed across various fields. Table 5 shows the distribution of book samples across the highest digit of the NDC (Nippon Decimal Classification) system, the most widespread book classification system in Japan. It turned out that social science and literature are the two prevailing fields in this corpus. Similarly, Table 6 shows the distribution of magazine samples across the classification adopted in *Periodicals in Print in Japan*. More than 70% of magazines belong to the single category of 'general-interest' that includes magazines like *Bungei Shunjuu* and *Chuuou Kouron*.

Table 5. Distribution of book samples in the production sub-corpus across NDC.

NDC	RATIO[%]	N SAMPLE	N SUW (estimation)
0 General	3.37	425	975,000
1 Philosophy	5.35	675	1,548,000
2 History	8.86	1,117	2,562,000
3 Social Science	25.56	3,222	7,391,000
4 Natural Science	10.44	1,316	3,020,000
5 Engineering	9.51	1,199	2,750,000
6 Industry	4.52	570	1,308,000
7 Arts	6.71	846	1,941,000
8 Language	1.83	231	529,000
9 Literature	19.24	2,425	5,564,000
Not Classified	4.59	578	1,326,000
TOTAL	100.00	12,604	28,915,000

Table 6. Distribution of the magazine samples in the production sub-corpus

CATEGORY	RATIO[%]	N SAMPLE	N SUW (estimation)
General-interest	70.58	1,927	3,400,000
Education, Academy	8.35	228	402,000
Political, Economy	4.34	118	209,000
Industrial	1.05	29	51,000
Engineering	13.96	381	673,000
Medical	1.72	47	83,000
TOTAL	100.00	2,730	4,818,000

Table 7. Examples of the XML tags defined for the BCCWJ.

TAGS ABOUT	XML ELEMENT	MEANING
Character	ruby	Indication of the pronunciation of Kanji characters (Japanese '*furigana*')
	missingCharacter	Indicate that the character is out of the JIS X0213 character set.
	correction	Correction of errata in the original text
Document structure	sample	Indicate the range of a sample
	article	Unit of text written by a single author about a single theme.
	title	Title given to a unit of text.
	cluster	Range of text referred to by a title.
	abstract	Abstract of summary about the article.
	authorsData	Info about the person(s) or organization that wrote the document.
	figureBlock	Indicate trace of figures, picture, and captions.
	list	Indicate the elements in a list or table.
	noteBody	The body of footnote or endnote.
	paragraph	Paragraph as shown by indentation.
	sentence	Sentence as indicated by symbols like ".!?"
	quotation	Citation from different documents, or transcription of spoken utterances.
	verse	Indicate the range of verse or poetry.

4.7. Encoding and annotations

The whole specimen of the BCCWJ will be encoded using the Japanese character set defined as JIS X0213:2004 using the Unicode (UTF16LE) as the character code. XML will be used as the basis of information exchange.

4.7.1. Tags about text structure

As shown in Table 7, various XML tags were designed to represent character information and document structure information. Figure 4 shows an example of the application of XML elements <article>, <title>, and, <cluster> to a sample taken form a magazine. In this case, the whole text involved in the figure constitutes a single article. Each title in the article is indicated by rectangles in real line, and the clusters corresponding to each of the titles are indicated by parenthesis. Dotted and real parentheses stand respectively for clusters of higher and lower hierarchies. Note in this example, only one cluster of lower hierarchy is shown for a higher cluster for the sake of clarity in presentation.

Figure 4. Example of the tagging of document structure by use of XML elements <article>, <title>, and <cluster>

4.7.2. Linguistic annotation

In addition to the tags mentioned above, linguistic annotations will be provided for the corpus. At the very least, SUW-LUW two-way POS information is to be provided for the whole BCCWJ. Our goal in this field is to achieve the accuracy of 98% in the automatic POS analysis of SUW (current accuracy being about 95% in passing).

More linguistic annotations like dependency-structure tagging, phrase-structure tagging, discourse-structure tagging, and anaphora tagging are also planned by the researchers belonging to one of the research groups of the priority-area 'Japanese Corpus' project under the direction of Professor Yuji Matsumoto of NAIST. They are also trying to develop a scheme for the integration of various annotations.

Lastly, it is probable that some of the advanced annotations can not be applied for the whole corpus because it is difficult to automate the annotation. Our plan is to establish a special subset of the corpus to which the effort of annotations is concentrated, as was the case in the advanced linguistic annotation of the CSJ-Core (See section 3).

5. Conclusion

Compilation of a large-scale balanced corpus of contemporary written Japanese is underway, aiming at the final public release in the year of 2011. Combined use of the CSJ and BCCWJ will provide reliable infrastructure for Japanese corpus linguistics. It will also give stimulation to many applied research fields like language education, language planning, and language processing for example. Especially, it is the present author's conviction that the education of a mother tongue and the planning of a national language should be supported by a firm basis of data-oriented linguistics. Without its support, discussions about the mother tongue can easily be plunged into utter confusion.

6. Acknowledgements

The current author is grateful to my colleagues Takehiko Maruyama and Yoko Mabuchi who prepared most of the materials used in chapter 4.

7. References

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing*, 8, pp.243-257.

Furui, S., M. Nakamura, T. Ichiba, and K. Iwano 2005, Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese. *Speech Communication*, 47, pp.208-219.

Kennedy, C. 1998. *An Introduction to Corpus Linguistics*, Addison Wesley Longman,

London and New York.
Maekawa, K. 2005. Quantitative analysis of word-form variation using a spontaneous speech corpus, Proceedings of Corpus Linguistics 2005, Birmingham (http://www.corpus.bham.ac.uk/PCLC/).
Maekawa, K. 2006. Analysis of Language Variation Using a Large-Scale Corpus of Spontaneous Speech, Invited speech at the *International Symposium on Linguistic Patterns in Spontaneous Speech*. Taipei (http://www.lpss.sinica.edu.tw/_Invited_Speakers.htm).