Design of a Balanced Corpus of Contemporary Written Japanese

Kikuo Maekawa

Department of Language Research
The National Institute for Japanese Language, Tokyo, Japan
kikuo@kokken.go.jp

Abstract

Compilation of a 100 million word balanced corpus of written Japanese, known as BCCWJ, is underway. This paper describes the design issues of this corpus. The corpus consists of three component sub-corpora differing in the sampling population and sampling technique. Two of them are compiled using the standard technique of random sampling. The third sub-corpora contains various mini corpora designed for special research purposes; this last sub-corpora will include at least 5 million word mini corpus of internet text.
Index Terms: Balanced corpus, corpus design, sampling

1. Introduction

Linguistic study of the Japanese language lags considerably behind world's standard as long as modern corpus linguistics is concerned. It is widely recognized by those who work in the field that one of the fundamental problems in Japanese corpus linguistics is the lack of balanced, or reference, corpus; a corpus that represents wide range of text genres existing in the target language.

To fill this lag, National Institute for Japanese Language (NIJL, hereafter) has launched a new corpus compilation project in the spring of 2006. The aim of this paper consists in the presentation of basic design issues of the new corpus.

2. Background

2.1. *KOTONOHA* and BCCWJ

NIJL has a long-term corpus development initiative for the modern and contemporary Japanese, which is known as the *KOTONOHA* project [1]. *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*. As shown in this figure, there are two components 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. *Taiyo* was a general-interest magazine read by very wide range of readers. There is also *Corpus of Spontaneous Japanese* (CSJ) that records spontaneous monologue of the present day Japanese (recorded in 1999-2002).

There is a corpus-to-be-compiled locating at the upper right corner of the figure. This component of *KOTONOHA* 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.

2.2. Priority-Area Project

A group of researchers in the Department of Language Research of NIJL 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 of BCCWJ.

At the same time, the NIJL group, together with researchers outside the NIJL, has applied for a MEXT (Ministry of Education) grant-in-aid for priority area scientific research for the compilation and analysis of BCCWJ.

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 project was blessed by the MEXT's announcement of the acceptance of the priority area program 'Japanese Corpus', which is a five year project of 2006-2010 and the principal investigator being the present author [2]

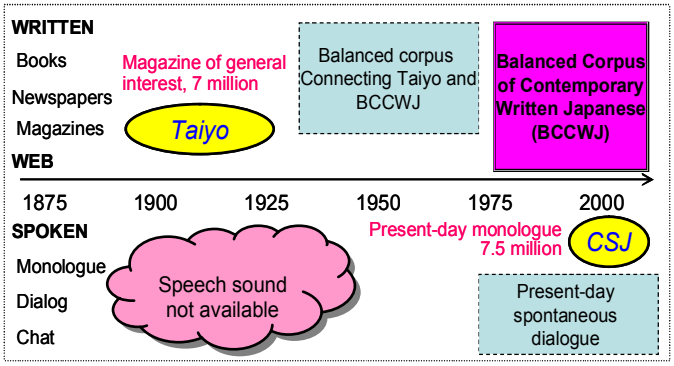

Figure 1. Schematic representation of KOTONOHA. Ellipse and box stand respectively for corpora publicly available and to be compiled. BCCWJ locates at the upper right corner. Dotted boxes are those to be compiled after the completion of the BCCWJ.

3. Structure of BCCWJ

One of the most important, and controversial, issues of the design of balanced corpus is that of securing representativeness (See [3] and [4] among many others). Needless to say, there is no 'one and the only one' solution to this issue; there are

multiple ways of securing corpus representativeness, depending on the purpose of corpora.

In the case of BCCWJ, we 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 sub-corpora. These sub-corpora will be treated separately in the following subsections.

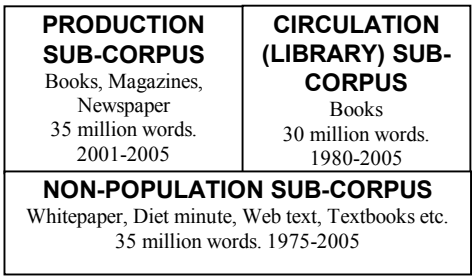

Figure 2. Three component sub-corpora of the BCCWJ

3.1. Production sub-corpus

The upper left-hand sub-corpus of Figure 2 is called 'production' sub-corpus. As suggested by the 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.

At this point, 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) and *Periodicals in Print in Japan*, for example. To be more exact, we estimated the total number of characters (or letters) 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.

Note especially that the composition ratio of genres (i.e. the ratio among samples of books, magazines, and newspapers) was determined on the basis of publication data. This makes crucial difference from corpora like the Brown Corpus and BNC, where the composition ratios of various genres were determined subjectively by specialists of English without making reference to objective data.

The total size of the sub-corpus is supposed to be about 34.7 million words (see section 3.5 for details), and according to the latest estimation, 74.1, 16.1, and 9.8% of the sub-corpus will be devoted to samples of books, magazines, and newspapers respectively.

3.2. Circulation sub-corpus

It is possible to conceive of balanced corpora that represent aspects of language usage other than the production aspect. Some users may find it important to obtain information about the reception of language. 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 accepted by multiple public libraries in the Tokyo Metropolis. The data was provided to us mostly by the courtesy of Tokyo Metropolitan Library.

It turned out that the total number of books registered in those libraries is about ten million, and they consist of 1,064,186 different books. So, 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 is supposed to be about 30 million words (see 3.5 below).

As for the time range, in principle, only those books to which ISBN (International Standard Book Number) is assigned are to be involved in the corpus. The population thus defined will cover roughly the last one quarter century, i.e. the years after 1980 when ISBN started to be adopted by Japanese publishers. This relatively wide coverage of time range makes important difference from the narrow time range of the production sub-corpus.

3.3. 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.

The sub-corpus contains mini corpora of **a)** White paper texts of about five million words, **b)** Internet text (Yahoo! Japan's internet bulletin board *Chiebukuro*) containing about five million words, and, **c)** Minutes of Japanese National Diet containing about five million words (covering both the Upper House, or *Sangi'in,* and Lower House, or *Shuugi'in*).

In addition to these, we plan to construct mini corpora of Japanese textbooks that will include texts of **d)** Textbooks used in the elementary, junior-high, and high-schools, and, **e)** Textbooks of Japanese as the second language.

It is important to note that these mini-corpora are linked to particular research activities in the NIJL and priority-area 'Japanese Corpus' project. like the compilation of textbook word frequency list and the comparison with the BCCWJ as a whole.

3.4. Specimen length

Two types of specimen texts differing in length will be taken from the sample books, magazines, and newspapers: fixed-length specimen of 1000-character long, and, specimen of variable length that covers minimal structural and meaningful unit of the sample. Most usually the latter specimen corresponds to the structural elements like section or chapter. According to the analysis of the pilot corpus, average length of variable-length specimen was 3900, 3000, and 1000 characters 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).

The variable-length specimen will be appreciated by the corpus users who are interested in the analysis of text- and discourse-structure.

On the other hand, we need fixed-length specimens in order to get statistical information as exact as possible about the statistical distribution of the population. It is especially important to know the exact estimation of the frequency of various Kanji characters (Chinese logograph). The frequency information is the very basis of the examination of the national Kanji lists for educational and/or daily usage purposes ('*Kyooiku*' and '*Jooyoo*' Kanji lists), which has been regarded to be one of the most urgent issues in the language planning of the Japanese language.

In samples taken from books, fixed-length specimens are usually included in the corresponding variable-length specimens. But in the case of newspapers samples, it is not usually the case, because variable-length specimens may become shorter than the corresponding fixed-length specimens. It is because most of newspaper articles are shorter than 1000 characters and in the case of newspaper samples variable-length specimens will not exceed an article. When it happens, the fixed-length specimen will take the rest of 1000 characters from 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 from the population. Fist we take fixed-length specimen, by choosing 1000 letters 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.

3.5. 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. The sizes of these corpora were determined as follows.

First of all, we started by determining the number of fixed-size specimen in the production sub-corpus required for various statistical inference purposes. Based upon our prior experience and computer simulation, the size was determined to be 10 million words.

Once this is fixed, the number of fixed-length specimen could 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 character length of a word (SUW, see 4.2) 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.

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 character length of a word.

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 1 for the result of estimation. See also section 3.4 for averaged length of variable-length specimen.

Table 1. Size of production sub-corpus

GENRE	RATIO[%]	N SAMPLE	SIZE (Word)
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

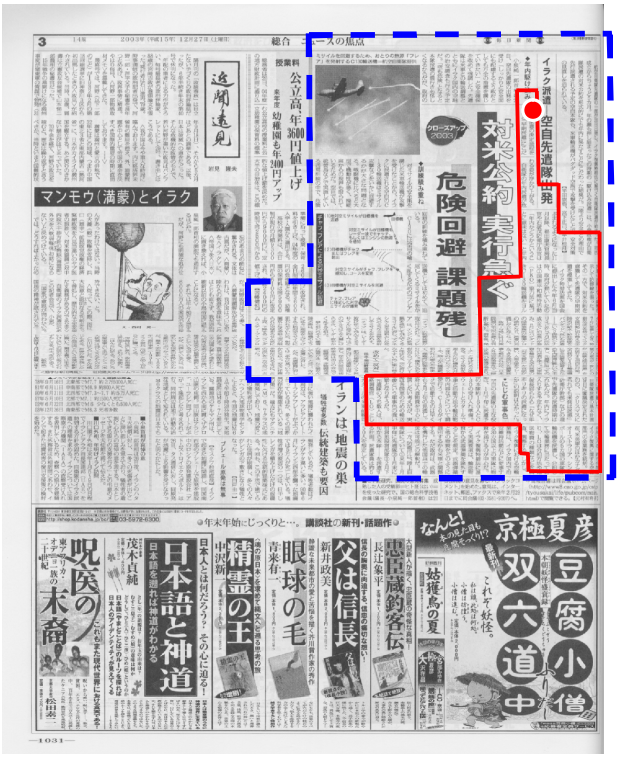

Figure 3. Example of fixed- and variable-length specimen in the case of a newspaper article. Letters used in figures and tables are not included in the specimen. Advertisements are also neglected.

As for the circulation sub-corpus, we have not finalized the design yet. A plausible possibility is to 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. But it has to be 35 million words at the very least to fulfill our official vow about the size of the BCCWJ.

3.6. Distribution of samples

Tables 2 and 3 show how the samples in production sub-corpus are distributed across various fields. Table 2 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. Social science and literature are the two prevailing fields in this corpus.

Table 2. Distribution of the book samples in the production sub-corpus across NDC numbers.

NDC	RATIO [%]	N SAMPLE	N WORD (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 3. Distribution of the magazine samples in the production sub-corpus

CATEGORY	RATIO [%]	N SAMPLE	N WORD (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

Similarly, Table 3 shows the distribution of magazine samples across the classification of magazines adopted in *Periodicals in Print in Japan*. It is interesting to see that more than 70% of magazines belong to the single category of 'general-interest' that includes famous magazines like *Bungei Shunjuu* and *Chuuou Kouron*.

4. Encoding and annotation

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.

4.1. Tags about text structure

Various XML tags were designed to represent character information and document structure information. The basic tag set about the character information includes <ruby>, <correction>, <missingCharacter>, <superscript>, <subscript> etc. The document structure tag set includes <sample>, <article>, <title>, <cluster>, <hierarchy>, <paragraph> and so on. Also, there are tags about the content of text like <abstract>, <list>, <noteBody>, <citation>, <speech> etc.

4.2. Linguistic annotation

In addition to the tags mentioned above, linguistic annotations are to be provided as well. They include POS information and dependency structure information among others.

For the POS annotation, two-way POS system of SUW (short unit word) and LUW (long unit word) will be adopted as in the CSJ. As for SUW, there are 16 POS categories at the top level of hierarchically organized POS system.

More linguistic annotations like phrase-structure, discourse-structure, and anaphora labeling are 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 [5].

5. Conclusion

The design and compilation of a large-scale balanced corpus of the present-day written Japanese is underway in the NIJL with the financial support from the MEXT priority-area program. The production and circulation sub-corpora of the BCCWJ will be world's first balanced corpus that was designed thoroughly on the basis of statistical sampling.

6. References

[1] http://www2.kokken.go.jp/kotonoha/

[2] http://www.tokuteicorpus.jp/

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

[4] Kennedy, C. *An Introduction to Corpus Linguistics,* Addison Wesley Longman, London and New York, 1998.

[5] http://www2.kokken.go.jp/csj/public/index.html

Acknowledgement

The author is very grateful for his colleague Takehiko Maruyama who kindly prepared most of the data about the current status of BCCWJ reported in section 3.