Balanced Corpus of Contemporary Written Japanese

Kikuo Maekawa
Dept. Language Research, National Institute for Japanese Language
10-2, Midori-cho, Tachikawa-shi, Tokyo 190-8561 JAPAN
kikuo@kokken.go.jp

Abstract

Construction of 100 million words balanced corpus of contemporary written Japanese is underway at the National Institute for Japanese Language. The unique property of the corpus consists in that the majority of its sample texts are selected randomly from well-defined statistical populations covering wide range of written texts.

1 Introduction

A serious problem in corpus-based analysis of the Japanese language is the lack of reliable balanced corpus. Most of the corpus-based studies of contemporary Japanese are based upon the analyses of text archive of newspaper articles, archive of copyright-expired literary work (*Aozora Bunko*), or crawling of the Internet text.

Putting aside the problems of the copyright-expired literary works, which are definitely too old to be the material for the study of the contemporary Japanese, lack of balanced corpus imposes two mutually related problems on linguistic studies. Most of newspaper articles are written by newspaper writers who are very much aware of the established writing style (and orthography) of newspaper articles. 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 are very much likely to include texts covering wide range of texts. 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 information about the genre of texts and/or the writers are usually missing. Moreover, the amount of retrieved texts can often be too large to be classified by hand.

There is also a problem of skewed distribution of texts caused by copyright protection. Copyright-protected materials, especially literary works, will not usually show up in the Internet.

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

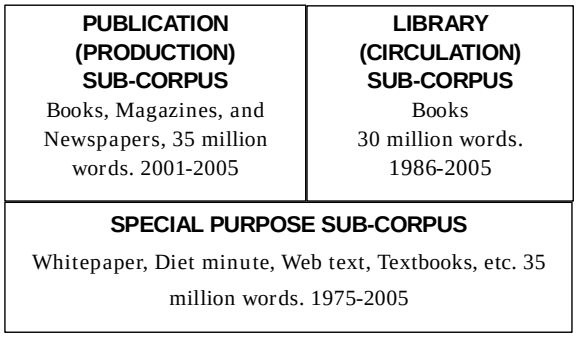

Figure 1. The three components of the BCCWJ.

2 Design of the BCCWJ

As shown in the figure 1, the BCCWJ consists of 3 component sub-corpora, viz., 'publication', 'library', and 'special purpose' sub-corpora.

2.1 Publication sub-corpus

The upper left-hand sub-corpus of the figure 1 is called 'publication' sub-corpus. This is also called

'production' sub-corpus. As the name suggests, 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 during 2001-2005.

The population was constructed using the sources that are publicly available; *J-BISC* (Japan Biblio Disc) and *Periodicals in Print in Japan* were used as the sources for books and magazines respectively. The data for newspapers was available from the association of newspaper companies (*Nihon Shinbun Kyoukai*). The total number of characters involved in the population was estimated and samples for the BCCWJ were drawn in the way that each character in the population had the same chance of being sampled.

It is to be noted at this point that the composition ratios among text 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 designs of corpora like the Brown Corpus and the BNC, where the composition ratios of various genres were determined subjectively by specialists of the English language without making reference to any objective data.

The total size of the sub-corpus is supposed to be about 34.7 million words; and, 74.1, 16.1, and 9.8% of the sub-corpus are to be devoted to the samples of books, magazines, and newspapers respectively.

2.2 Library sub-corpus

The second sub-corpus is called 'library' or 'circulation' sub-corpus. The sampling population for this sub-corpus was the whole books registered in at least 13 public libraries in the Tokyo Metropolis. The population thus defined contains about 335,000 books. According to our estimation, more than 48 billion characters are included in this population, which is nearly the same amount as the population of the book part of the publication sub-corpus.

There are two important differences between the publication and library sub-corpora. Firstly, the texts of the library sub-corpora represent those books that were accepted by a certain number of readers, while the texts in the publication sub-corpora have no guarantee of the sort.

Secondly, the library sub-corpus covers the period of time 1986-2005 (1986 was the year when the ISBN book classification system was adopted by most of major publishing companies), while the period of time covered by the publication sub-corpus is 2001-2005.

2.3 Special-purpose sub-corpus

The last sub-corpus is called 'special purpose' or 'out-of-population' sub-corpus. This is the aggregate of various special purpose mini corpora, and, unlike the former two sub-corpora, some of the mini corpora are not sampled using the technique of random sampling (because the populations can not be defined).

The mini corpora currently include texts of governmental white papers, Internet text (Yahoo! Japan's bulletin board *Chiebukuro*), minutes of the national diet, school textbooks, and best-selling books of the past 30 years. Laws and academic papers will also be included.

Most of these mini corpora contain about 5 million words, and will be utilized in the language policy oriented activities of the NIJL, including the revision of the National list of Chinese Characters for Daily Usage (*Jouyou kanji hyou*).

3 Funds and the current status

The compilation of the BCCWJ is supported by the budget of the NIJL and the MEXT (ministry of education) Grant-in-Aid for Scientific Research Priority Area Program "*Japanese Corpus*" (2006-2010) [1].

As of September 2007, texts containing about 30 million words have been sampled and stored in the NIJL server, and, the full-text query of the 10 million words texts that are copyright cleared are publicly available on the web [2].

Upon its completion, the BCCWJ will be the world's first balanced corpus that is designed and compiled based upon rigid statistical sampling. This will open up a new possibility in Japanese linguistics, and the design of language corpora in general.

References

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

[2] http://www.kotonoha.gr.jp/demo/