KOTONOHA: A Corpus Compilation Initiative at the National Institute for Japanese Language

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

Today's natural language and speech processing technology depends heavily upon statistical approach; hence there is need for large-scale language corpora. As long as the Japanese language is concerned, it was not linguists but engineers that played a leading role for the development of language and speech corpora in the early stage of corpus-related studies.

In the last half of the 1990s, however, linguists became more and more interested in the corpus-bases analysis of the Japanese language. But it turned out quickly that the characteristics of corpora needed in engineering and linguistics did not necessarily coincide. Most obvious difference consisted in the temporal coverage of corpora; while engineers dealt with the data of contemporary language, linguists often had a need for the data of the past language. Also obvious was the difference in the size and quality of data: in many cases, engineers put more emphasis on the size rather than the quality of data. Lastly, it was also the case that, generally speaking, linguists were interested in much wider range of language varieties than engineers.

The aim of the KOTONOHA initiative of the National Institute for Japanese Language (NIJL) consists in providing a series of corpora that cover the whole range of the Japanese language. KOTONOHA (ancient Japanese meaning 'language') is a cover term for a series of component corpora covering both written and spoken, and, contemporary as well as historical varieties. The component corpora are designed primarily for linguistic research, but the possibility of engineering application is not excluded from the corpus design, especially when a corpus is concerned with contemporary spoken Japanese.

So far, there are 2 corpora of KOTONOHA that have already been made publicly available. The *Taiyô Corpus* deals with the written Japanese of at around the boundary of the 19th and 20th centuries; it is the time when the grammar of the written Japanese changed drastically. The *Corpus of Spontaneous Japanese* (CSJ) is a spoken language corpus designed for the statistical learning of the language- and acoustic-models for the next generation automatic speech recognition system that is capable of handling more-or-less spontaneous speech. Although the CSJ was built primarily for engineering application, a subset of the corpus was richly annotated for the phonetic and linguistic study of spontaneous speech. The annotation included X-JToBI labels for segmental and prosodic features, clause boundary labels, and, the labels of topic boundary.

Since 2006, the corpus compilation group of the NIJL devotes concentrated efforts to the compilation of the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ). It is a long-awaited balanced corpus of contemporary Japanese and probably the most important corpus of the whole KOTONOHA initiative. Among the 100 million words samples of the BCCWJ, about 35% is chosen by means of random sampling from the populations covering the books, magazines, and newspapers published in the years 2001-2005. In addition, 30% of the samples are randomly chosen from the population of the books registered in the public libraries of Tokyo Metropolis and published during the years 1985-2005. And, the resulting 35% is devoted to various mini corpora including the texts of governmental white papers, minute of the National Diet, textbooks of elementary and secondary education, bestselling books, the Internet texts of blog and bulletin board, and so forth.

Although the compilation of the BCCWJ as a whole is not an easy task, the biggest problem lies in the clearance of copyright protected materials. For example, as long as the book samples are concerned, more than 10,000 permissions from the copyright holders have been obtained during the last 30 months, but they cover only 40% of the totality of 24,000 book texts that have been sampled so far. This is not because the copyright holders were unwilling to give permission. More than 90% of copyright holders gave us permission, once we could make contact with them. The problem lies in the fact that it is often impossible to make contact with the copyright holders.

Though we will continue working hard to obtain permissions from the copyright holders, an amendment of Japanese copyright law by which ‘fair-use’ of copyright protected materials becomes possible is strongly needed. Fortunately, the Copyright Division of the Agency of Cultural Affairs is now preparing a bill for such amendment. At the earliest, the deliberation of the amendment bill will be started sometime in 2009.

Anyway, corpus materials corresponding to about 70 million words have already been sampled and stored in a server, and about 40 million words copyright-cleared texts are available for full-text search in a demonstration web site (http://www.kotonoha.gr.jp/demo/). The BCCWJ will be publicly available in 2011.