Compiling a Balanced Corpus of Modern Japanese: Design Issues and Implications for Japanese Lexicography

Yasuo MORITA, Takahiro NAKAMURA, Hiroshi AIZAWA
Shogakukan Inc.
2-3-1 Hitotsubashi Chiyoda-ku,
Tokyo 101-0081
JAPAN
June TATENO
NetAdvance Inc.
2-30 Kanda-Jinbocho Chiyoda-ku,
Tokyo 101-0051
JAPAN
Yukio TONO
Department of Foreign Language and Culture
Meikai University
8 Akemi, Urayasu,
Chiba 279-8550
JAPAN

Abstract
An extensive, balanced corpus of the Japanese language comparable in size and structure to the British National Corpus (BNC) has so far not been compiled. Consequently, the practice of compiling general-purpose monolingual dictionaries on the basis of corpora has not yet been fully established in our country. Shogakukan, one of the leading publishers in Japan, has launched a project to compile a balanced Japanese-language corpus in order to produce new monolingual dictionaries of contemporary Japanese. Shogakukan has a long history and has been producing over 2,000 titles a year, ranging from general fiction, non-fiction books, magazines, and comics to dictionaries. These in-house materials are of sufficient variety and quantity to give us a headstart in producing a large-scale, balanced corpus of the Japanese language, the *Shogakukan Contemporary Japanese Corpus*, with the collaboration of academic research institutions and other publishing companies. In this paper we will focus on the following two points: (1) how to define the concept of balance in terms of genre or sub-corpora proportions by using breakdown statistics for publications in the domestic market in Japan; (2) how to estimate the optimal corpus size (i.e., the minimal amount of corpus data that can still adequately represent all the different linguistic behaviors of the candidate dictionary entries.).

1 Introduction

In Japan, the use of empirical corpus data for lexicographic purposes has yet to become established. Whilst other countries such as Korea, Poland and Spain have already launched national or government-funded projects to compile so-called "British National Corpus (BNC) counterparts" for their own languages, no nation-wide project to create a corpus of

Japanese has yet been planned in Japan. Japanese lexicographers still primarily rely on citation databases for the compilation of large-scale general-purpose monolingual dictionaries. The *Nihon Kokugo Dai-Jiten* (The *Unabridged Dictionary of Japanese* by Shogakukan), for example, followed the historical principles of the *Oxford English Dictionary*. 750 thousand illustrative examples were taken from a database of 2.5 million citations. In order to prepare this citation database, 150 researchers spent about 10 years selecting examples from documents written in different eras. However, we should point out that this is a rather rare case and that intermediate-sized dictionary projects have usually relied on much smaller citation databases and on the intuition and experience of their lexicographers. Projects such as the *Nihon Kokugo Dai-Jiten* will, however, sooner or later have to modernize and computerize their practices and update their methodology, and start amassing appropriate language data for compiling a dictionary along corpus-based lines.

Having recognized current trends in the field, Shogakukan has decided to develop a large-scale balanced Japanese corpus, the first of its kind in Japan, for the purpose of corpus-based Japanese dictionary-making. We are aiming to release the final version of this corpus, the *Shogakukan Contemporary Japanese Corpus*, in 2006. Shogakukan, which has a long history as a general publisher, produces over 2,000 titles a year across a range of genres that includes general fiction and non-fiction books, magazines and comics in addition to dictionaries. We will construct the Contemporary Japanese Corpus in the first phase (year 2003-2006) by collaborating with other publishing companies and researchers, and then plan to add spoken and historical corpora as sub-corpora in the next phase (2007-2010). In this paper, we will first give an overview of the design of our corpus, focusing on the configuration of sub-corpora to be included and the issue of corpus size and then discuss how the corpus may be exploited for compiling an empirically-based dictionary of Japanese.

2 Overview of the Contemporary Japanese Corpus

There is a huge gap in written Japanese in terms of grammar and orthography for documents produced before the 20th century and those after, and it is technically very hard to process by machine all the written Japanese from the 8th to the 21st century. Therefore, we have decided to limit our scope to "contemporary" Japanese only.

Our Corpus design follows that of the BNC, especially in terms of the target selection features (domain, time and medium, etc), descriptive features for written texts (author details, target audience and place of publication, etc.) and the proportion of written and spoken texts. As regards the size of the corpus and target domains, we will make a decision based on the number of publications inside Japan and the size of Japanese dictionaries that we will produce in the future.

3 Configuration of sub-corpora

In defining the content of the corpus, we used the publication statistics in the domestic market in Japan as our main guide. The data were based on *Books in Print in Japan* by Shuppan News Co., which annually publishes an excellent resource in the form of a general directory of books published in Japan. The editions published after 1972 have been particularly valuable as they contain breakdowns by domain of the number of publications and sales, among other figures. Table 1 shows the number of books published in each of the

domains reported in *Books in Print in Japan* over the period of thirty years from 1972 to 2001. (Note that these figures do not include periodicals such as newspapers and magazines. Note also that the figures *do* include re-publications of older texts, such as new editions of 19th-century novels or collections of literary classics. These raw figures will therefore require a certain degree of adjustment.).

Domain	No. of books	Percentages
General	45,899	4.19%
Literature	234,205	21.40%
Natural science	93,225	8.52%
Technology	106,094	9.69%
Social science	263,551	24.08%
History	80,124	7.32%
Industry	49,861	4.56%
Arts	137,359	12.55%
Language	26,852	2.45%
Philosophy	57,373	5.24%
Total	1,094,543	100.00%

Table 1: New Publications in Japan by Genre, 1972-2001

We will structure the sub-corpora of *the Contemporary Japanese Corpus* broadly in these proportions, while taking into account the genre divisions used in the BNC.

4 Estimating the corpus size relative to the number of entries

4.1 The case of English corpora

Although it is true that for some corpus research purposes the dictum "the bigger, the better" is true, the larger the size of a corpus the more the administrative work involved in copyright clearance, and the greater the costs in terms of usage rights fees (where applicable). It is therefore desirable to have some rough estimate of the optimal corpus size in relation to the number of lexicographic entries the corpus will be serving as a database for. In our paper, we will first show that Zipf's Law, which states that frequency (f) multiplied by rank (r) will tend to produce a constant k, particularly in the mid-frequency range, is not a very good predictor of corpus size in the case of low-frequency items. In other words, it is observed that as one increases the size of the corpus, words of medium frequency and above exhibit a linear increase in occurrences, while words of very low frequency do not increase as much and the supply of new words dries up. We will demonstrate this by comparing data from some existing English corpora (in particular, the BNC and the Brown Corpus).

4.2 The case of Japanese corpora

Since Japanese is very different from English, we need to examine the effects of the characteristics of the Japanese language on the results of the corpus size estimation. Since we do not have explicit word boundaries in text, tokenization as well as morphological analysis are much more complicated than they are for English. Moreover, compared with English, Japanese has a relatively high incidence of function words, corresponding to prepositions and particles in English. Hence if we simply count the morphologically analyzed units, there is a danger that we may overestimate the size of the corpus needed, which may in turn result in less efficient retrieval operations.

The following example is from *A Scandal in Bohemia* by Conan Doyle. Content words make up 60% of the words in the English original but 50% in the Japanese translation.

My own complete happiness, and the home-centred interests which rise up around the man who first finds himself master of his own establishment, were sufficient to absorb all my attention, while Holmes, who loathed every form of society with his whole Bohemian soul, remained in our lodgings in Baker Street, buried among his old books, and alternating from week to week between cocaine and ambition, the drowsiness of the drug, and the fierce energy of his own keen nature.

(-- *A SCANDAL IN BOHEMIA*□*Conan Doyle)*

Watasi ha **arawasi** you no nai **siawase** wo **te** ni **si**,**hajimete ikka** no **syujin** to **natta mono** no **rei** ni **more** zu, **mi** no **mawari** ni **aru katei** wo **tyusin** to **sita seikatu** no **omosirosa** wo **miidasi** ,**subete** no **kansin** wo **mukete** ita. Ippou **Homuzu** ha **aimokawarazu** sono **bohemiankisitu** yueni **jiyuuhonpou** de , **syakougirei** nado **kiratte hitodukiai** wo **issai sake**, **izentosite Beka-gai** no warera ga **gesyuku** ni **hikikomori** , **kosyo** no **naka** ni **kao** wo **umete** , **kokain** to **koumyousin** , tumari **mayaku** ni **huketta** ri , umaremotta **tensei** no **surudosa** de **nessin** ni **sigoto** ni **bottou** suru **koto** wo **kougoni kurikaesite** ita no datta

(Contents words in bold)

We calculated the relative percentages of content words and function words in the English language as represented by the BNC data and established that the content words typically account for about 49 % on average. We also examined the typical proportion of content words in Japanese using our in-house data. The results are shown in Table 2:

Domain: Imaginative

Files	115
Total size (bytes)	57,662,224
Average	501,410

Number of word	total	%	unique
Content word	7,424,633	39.96	109,905
Functional word	8,271,428	44.51	308
Punctuation and etc.	2,885,583	15.53	101
Total	18,581,644	100.0	110,314

Table 2: Rough estimate of the ratio of content to function words in Japanese texts (imaginative)

This table shows the results from our analysis of 115 files composed of Japanese novels. The results show that the proportion of content words is about 40% of the overall number of words. We will investigate whether or not the collections of texts from other domains or genres will exhibit any significant difference from this figure. Calculating these ratios helps us in estimating the ideal corpus size, since content words rather than the total number of words should be used as the basis for comparing a Japanese corpus with English corpora such as the BNC, and for calculating the minimal size of corpus which will provide sufficient corpus evidence for any given number of entries.

5 Using a contemporary Japanese corpus to compile a monolingual Japanese dictionary

As we mentioned in the Introduction, we have had the experience of engaging in a large dictionary compilation project comparable to the *Oxford English Dictionary*. It was an extremely difficult task to compile the dictionary based on millions of citation cards. Nevertheless this experience has benefited the lexicographers in the company in terms of several key transferable skills which are necessary in the compilation and editing of any dictionary, such as (a) categorizing the cards in an alphabetical or thematic order, (b) making a careful selection of cards by checking the sample sentences in light of their definitions, (c) revising definitions of words based on the citation cards, (d) adding new definitions of words, (e) confirming the wording of the sample sentences by going back to the original source, and (f) selecting new words to be added to the dictionary. Although these processes were all done manually by hand in the last project, the skills and experience that lexicographers acquired throughout the dictionary-making process in such a traditional context should also help in the upcoming corpus-based lexicographic production process. In particular, the process of extracting the necessary information for a dictionary from a huge number of citation cards by careful reading and examination of the contexts will not be very different from the process of accessing corpus data and extracting lexicographically useful information from them.

Through this project of compiling the Contemporary Japanese Corpus, we aim to create a new work paradigm for Japanese lexicography, in which professional lexicographers rely on computer corpus data for writing and revising dictionary entries Shogakukan has developed two kinds of user-friendly software that are easy for lexicographers to make use of without any special knowledge of text processing. One is the Shogakukan Corpus Query System, with which one can search for words in large corpora on-line in-house as well as from outside the company, and the other is the Shogakukan Language Toolbox, which helps lexicographers to select necessary data (e.g. verb subcategorization data) from large corpora by batch processing. The former has been used for indexing the BNC for the Corpus Search Service which was launched in August 2003 as the " Shogakukan Corpus Network" (http://www.corpora.jp) and has received very positive evaluations. In the presentation, we will demonstrate some of the functions of these tools. Both services are currently being adapted for use with Japanese language texts. When this is done, dictionary-making in Japan will become a combination of a high-performance database and excellent lexicographers. Through this combination of sophisticated corpus-based tools and experienced lexicographers, Shogakukan are confident that we will soon bring to fruition the first corpus-based dictionary of the Japanese language.

6 Conclusion

Whilst Japan has already had expertise in Natural Language Processing and Information Retrieval, Japanese corpus-based lexicography is still in its infancy primarily because there has not been much interest by the Japanese government in creating a large dictionary or encyclopedia as a national project. Shogakukan hopes to take lead in this area and provide people in Japan with an innovative, truly corpus-based dictionary. Our current task is to produce a 3-million word XML or DB format corpus by the end of 2004, tagged with parts of speech. In 2005, we aim to obtain 10 million words. By 2006, we hope to acquire a 30 to 50 million-word balanced corpus.

7 References

Manning, C. D and H. Schutze, 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: MIT Press