Evaluation of Japanese Web-based Reference Corpora: Effects of Seed Selection and Time Interval

Motoko Ueyama

1 Introduction

The World Wide Web is an enormous resource of accessible textual documents, and there is by now a considerable amount of work on using the Web as a source of linguistic data for a variety of linguistic and language technology tasks (see, e.g., the papers collected in Kilgarriff and Grefenstette 2003). A promising approach to the use of the Web for linguistic research is to build corpora by running automated queries to search engines, retrieving and post-processing the pages found in this way (e.g., Ghani et al. 2003; Baroni and Bernardini 2004; Sharoff this volume). This approach differs from the traditional method of corpus construction, where one needs to spend considerable time finding and selecting the texts to be included, but can have perfect control over contents. With the aforementioned automated methods, the situation is reversed: one can build a corpus in very little time, but without good control over what kinds of texts are included in the corpus. These automated methods, despite the almost complete absence of quality control, have made it possible to construct written corpora for linguistic research in a quick and economic manner. This is good news for researchers who urgently need large-scale balanced corpora (i.e., something equivalent to the British National Corpus) for the language of their interest, but who have no access to such corpora. This is the case for researchers working on the majority of the world's languages, including Japanese (see Goto 2003 for a survey of Japanese corpora currently available for research purposes).

The pioneering work in the automatic construction of Web corpora has been done by the CorpusBuilder project (see, e.g., Ghani et al. 2003) that developed a number of related techniques to build corpora for languages with fewer NLP resources. Ghani and colleagues evaluated the relative performance of their proposed methods in terms of quantity of retrieved pages. However, they did not provide a qualitative assessment of their corpora, such as a classification of the pages. Baroni and Bernardini (2004) introduced the BootCaT tools, a free suite of Perl scripts for the automated, possibly iterative construction of corpora via Google queries. While the tools were originally intended for the development of specialized language corpora and terminology extraction, they can also be used to construct general-purpose corpora by selecting appropriate query terms. The BootCaT tools were used for this purpose by Baroni and Ueyama (2004), Ueyama and Baroni (2005), Sharoff (this volume).

As mentioned earlier, Japanese is one of the languages for which general balanced corpora are not available. In the aforementioned studies (Baroni and Ueyama 2004; Ueyama and Baroni 2005), we built two Japanese Web corpora with the BootCaT procedure. In this study, we build another Japanese Web corpus with the same procedure, and conduct an evaluation by comparing the newly built corpus with our two other Japanese corpora and Sharoff's corpora.

Although a considerable amount of work has been done on ways to use the Web as a source of linguistic data, there are only few studies that have evaluated Web corpora, see e.g., for qualitative analyses, Fletcher (2004), Sharoff (this volume), Ueyama and Baroni (2005). Fletcher (2004) constructed a corpus of English via automated queries to the AltaVista engine for the 10 top frequency words from the British National Corpus (henceforth BNC) and applied various post-processing steps to reduce the "noise" in the data (duplicates, boilerplate, etc.). He compared the frequency of various n-grams in the Web-derived corpus and in the BNC, finding the Web corpus to be 1) more oriented towards the US than the UK in terms of institutions, place names and spelling; 2) characterized by a more interactive style (frequent use of first and second person, present and future tense); 3) permeated by information technology terms; 4) more varied (despite the fact that the Web corpus is considerably smaller than the BNC, none of the most common 5,000 words in the BNC were absent from the Web corpus, but not vice versa). Properties 2) and 4) challenge the view that Web data are less fit to linguistic research than a carefully balanced corpus of texts obtained in other ways.

Sharoff (this volume) uses an adapted version of the BootCaT tools to build Web-derived corpora for English, Russian and German. The corpora are constructed via automated Google queries for random combinations of frequent words extracted from existing corpora. He classifies 200 documents randomly selected from each corpus in terms of various characteristics, including the topic domains of each document, analyzed using the BNC classification system (with some adaptations). He finds that, in a comparison with the BNC, the English Web corpus is richer in exemplars belonging to the technical and applied science domains. He also compares word frequencies his Web corpora with reference corpora in English and Russian, and newswire corpora in English, Russian and German. His results show that the Web corpora are closer to the reference corpora than to the newswire corpora, also confirming Fletcher's findings about the Web being characterized by a more interactive style and more lexical variety.

In an already mentioned previous study (Ueyama and Baroni 2005), we qualitatively evaluated two Japanese Web corpora built in 2004 and 2005 with the use of the BootCaT tools. These are the corpora that here we call Genki 2004 and Genki 2005: see section 2 for details. The analysis showed that both corpora contained many documents produced by non-professional writers, characterized by everyday life topics and by an often informal, spontaneous, interactive style. Compared to Sharoff's results, we see that this text type is more dominant in our Japanese corpora than in any of his corpora in English, German, and Russian. We suspect that this difference between Sharoff's corpora and ours, i.e., a higher proportion of personal, spontaneous, interactive text in the latter, may be due to differences in seed choice. Our seeds, having been extracted from a basic vocabulary list from a Japanese textbook, are more often related to everyday life domains. In contrast, Sharoff's seeds are picked from existing traditional corpora (e.g., the BNC), and thus they tend to reflect some of the domains well represented in these corpora that are also common on the Web.[1]

The difference between Sharoff's and our results leads us to ask how different seed selection strategies affect the nature of resulting Web-based corpora. This is investigated by Ciaramita and Baroni (this volume) in a quantitative way. In this study, we perform a qualitative investigation, building and analyzing Japanese Web corpora using as seeds both words from a basic Japanese vocabulary list and words from Sharoff's English word list (based on the BNC) translated into Japanese. We conduct a relatively in-depth evaluation of the two resulting corpora in terms of domains, genres and typical lexical items, and discuss our findings in an attempt to answer the research question just described.

Another essential factor that affects Web corpus construction is time interval. It is well known that search engine indexing contin-

[1] A difference in the nature of the English and Japanese Webs, however, should not be completely ruled out, given a recent survey that indicates that the absolute number of blogs in Japanese is higher than the number of blogs in English. See `http://www.sifry.com/alerts/archives/000433.html`

uously changes, which is expected to strongly affect query results, and, consequently, the resulting Web corpus. The second goal of the study is therefore to investigate the effect of time interval and attempt to tackle the important issue of how “stable” the results of search engine queries are over time. For this purpose, we compare two Japanese Web corpora that we built at 10 months’ distance from each other (in July 2004 and April 2005, respectively) with the use of exactly the same automated procedure and seeds. As for the investigation of the effects of seed selection, we analyze the distributions of domains, genres and typical lexical items in each corpus.

The rest of the paper is structured as follows. In section 2, we present the procedure used to build our three Japanese Web-based corpora (Genki 2004, Genki 2005, BNC-seeded 2005) and describe the characteristics of each corpus briefly. In section 3, we describe our corpus classification methods and present our results, while section 4 presents the evaluation of typical lexical items for each of the three corpora. Finally, in section 5 we discuss our findings and conclude by suggesting directions for further study.

2 Corpus construction

In this section, we describe our three Japanese Web corpora. We built the first two corpora with the same automated procedure and seed terms, but at two different times: the Genki 2004 corpus in July 2004, and the Genki 2005 in April 2005 (these were the corpora analyzed in Ueyama and Baroni 2005). The BNC-seeded 2005 corpus was built in August 2005, using the same procedure but different seeds.

For the Genki 2004 and 2005 corpora, in order to look for pages that were reasonably varied and not excessively technical, we considered that we should query a search engine (Google in our case) for words belonging to the basic Japanese vocabulary. Thus, we randomly picked 100 words from the word list of *Genki*, an elementary Japanese Textbook (Banno et al. 1999; hence the name of the corpora): e.g., *tenki* "weather", *asagohan* "breakfast", *suupaa* "supermarket", *tsumetai* "cold". For the BNC-seeded 2005 corpus, we randomly picked 100 words from the list of 500 query terms that Sharoff extracted from the BNC to build his English Web corpus,[2] and translated those words into Japanese. The seeds that were selected for the construction of the BNC-seeded 2005 corpus vary more greatly in terms of domains (that include society, politics, history, computer technology) than the ones used for the two Genki corpora, that are very basic. We coherently translated the dictionary form of English verbs and adjectives into the dictionary form of their Japanese equivalents, although it is possible in theory to choose non-dictionary forms for Japanese translation candidates (e.g., formal present tense forms). In case both non-loanword and loanword varieties are available in Japanese, we employed the one that seems to be more common, which was expected to help to increase query hits: e.g., we translated "pattern" into *pataan* (loanword alternative), not *mohan* or *kata* (non-loanword alternatives).

All three Japanese Web corpora were built using the BootCaT tools mentioned earlier (Baroni and Bernardini 2004). We randomly combined the 100 seed terms into 100 triplets, and we used each triplet for an automated query to Google via the Google APIs (`http://www.google.com/apis`). The rationale for combining the words was that in this way we were more likely to find pages that contained connected text (since they contained at least 3 content-rich words). We used the very same triplets both in July 2004 and in April 2005 (for the Genki 2004 and 2005 corpora, respectively), while we created and used a new set of 100 triplets in August 2005 (for the BNC-seeded 2005 corpus). For each query, we retrieved maximally 10 URLs from Google, and we discarded

[2] `http://corpus.leeds.ac.uk/internet/seeds-en`

duplicate URLs. This gave us a total of 894 unique URLs in June 2004, 993 in April 2005, and 908 URLs in August 2005. Notice that, while for the purposes of our qualitative evaluation we are satisfied with corpora of these sizes, the same procedure could be used to build much larger corpora.

We compared the Genki 2004 and 2005 corpora in order to find how many URLs are present in both corpora. Interestingly, only 187 URLs were found in both, leaving 707 URLs that were retrieved in the Genki 2004 only and 806 URLs that were retrieved in the Genki 2005 only. Thus, with respect to the Genki 2005 URL list, the overlap with the previous year is of less than 20%. Moreover, there is of course no guarantee that the webpages corresponding to overlapping URLs between the two corpora did not change in terms of contents. To quickly investigate this point, we randomly selected 20 out of the 187 URLs retrieved in both years, and compared the 2004 and 2005 texts. We found that the two versions were identical in terms of contents for only 13 of the 20 URLs (65%), while the remaining pages had been modified (mostly for content updates). The changes in retrieved pages raise the question of whether the retrieved corpora are also different in terms of the nature of their contents or whether they are essentially comparable. This question will be examined later in section 3, on the basis of the results of the genre classification analysis. The overlap of URLs decreases even more between the Genki 2004 and BNC-seeded 2005 corpora. Only 11 URLs were present in both corpora. With respect to the Genki 2005 URL list, the overlap of URLs is only 1%.

For each URL, we (automatically) retrieved the corresponding webpage and formatted it as text by stripping off the HTML tags and other "boilerplate" (using Perl's `HTML::TreeBuilder` module and simple regular expressions). Since Japanese pages can be in different character sets (in particular, shift-jis, euc-jp, iso-2022-jp, utf-8), our script extracts the character set in which a page is

	total documents	total tokens	average size	error rate
Genki 2004	894	3,473,451	3,885	5%
Genki 2005	993	4,468,689	4,500	6%
BNC-seeded 2005	908	5,732,080	6,313	5%

Table 1. Total documents, total tokens, average size of tokens per document, and error rate in the Genki 2004, Genki 2005, and BNC-seeded 2005 corpora

encoded from the HTML code, and converts from that character set into utf-8. Since Japanese text does not use white space to separate words and characters, we used the ChaSen tool (Matsumoto at al 2000) to tokenize the downloaded corpora. However, ChaSen expects input and output to be coded in euc-jp, while our text-processing scripts are designed to receive text input coded in utf-8. To solve the problem of coding incompatibility, we used the `recode` tool[3] to convert back and forth between utf-8 and euc-jp.

According to the results of the ChaSen tokenization, the Genki 2004 corpus contains 3,473,451 tokens (about 3.5M); the Genki 2005 corpus 4,468,689 tokens (about 4.5M); the BNC-seeded 2005 corpus 5,732,080 tokens (about 5.7M).

Comparing the two Genki corpora, we have noticed that in Genki 2005 not only did the repeated queries find more and different URLs – they also found URLs that contained more text. This is illustrated by the average document size summarized in table 1. The BNC-seeded 2005 corpus, in turn, shows an increase of the total tokens of about 27%, and an increase of average document size of about 40% with respect to the Genki 2005 corpus, although the total document count decreases. We discuss the issue of the apparent trend of increase in corpus size and average document size in section 3, where the results of the corpus classification analysis are presented. We found (manually) that some pages did not contain any substantial amount of text: e.g., the ones that were not

[3] `http://recode.progiciels-bpi.ca/`

decoded properly, the ones that contained a warning message only, duplicates that were not removed, and so on. The ratio of these types of pages was approximately 5% for all the three corpora. We consider that this error rate is tolerable in the sense that the wide majority of text is usable.

3 Corpus classification

For the qualitative evaluation of our Japanese Web corpora, we manually classified all 894 pages of the Genki 2004 corpus, and 300 randomly selected pages each from the Genki 2005 and BNC-seeded 2005 corpora, in terms of topic domains and genre types.

3.1 Classification systems

3.1.1 Domains

For the classification of webpage domains, we adopted the classification system proposed in Sharoff (this volume), so that our results are directly comparable to his. We used the following nine categories:

natsci agriculture, astronomy, meteorology, ...

appsci computing, engineering, medicine, transport, ...

socsci law, history, sociology, language, education, religion...

politics

business e-commerce pages, company homepages, ...

life general topics related to everyday life typically for fiction, diaries, essays, etc...

arts literature, visual arts, performing arts, ...

leisure sports, travel, entertainment, fashion, hobbies ...

error encoding errors, duplicates, pages with a warning message only, empty pages

If a topic seemed to belong to more than one domain, we just selected one trying to be coherent. For example, we classified the webpages dedicated to a specific personal interest into the leisure domain, although the personal interests themselves are often related to everyday life, which is classified as the life domain (e.g., cooking, pets, etc.).

3.1.2 Genres

Webpages contain various genre types, including some attested in traditional corpora, e.g., news and diaries, and some newly emerging in Internet use, e.g., blogs (see Santini 2005). The situation is complicated by the fact that some documents can be a mix of more than one genre type (e.g., news report with an interactive discussion forum). Under these circumstances, it is not a simple task to classify Web documents by genre types. For the current study, the author first went through a good amount of the webpages to get a general idea of the distribution of genre types, and then selected the following 27 genre types as the final set:

blog personal pages created by users registered at blog servers that provide a ready-made page structure that, typically, include a diary with a comment section

BBS bulletin board sites; interactive discussion pages where multiple users can exchange messages with a topic-comments structure

diary a good example of an "adaptive" genre type that also exists in traditional written texts (see Santini 2005)

personal personal homepages not created through a blog service; less interactive than blogs since there is no interactive comment section

argessay essays written in an argumentative rhetoric style that present opinions, typically, on political or social issues

essay pages that state personal experiences, interests, feelings in a non-argumentative manner

novel another example of an adaptive genre type

commerinfo pages that present information to promote services or sell products

instrinfo pages designed to help readers to perform a certain task (how-to guides, guidelines, tips...)

info pages that present information that pertain to initiatives, events, resources and projects related to a certain topic without commercial or educational purposes (e.g., time/place of an upcoming event, political party manifestos, introduction to some academic program. . .)

teaching materials for instruction, typically, language teaching (e.g., example sentences, language exercises, ...)

news journalistic news; another adaptive genre type

njnews non-journalistic news, such as community pages

magazine Web magazine

acreport reports of academic research

report reports that present contents that pertain to a certain topic

review product/service evaluation, critique of arts, music, literature, etc.

comments comments directly sent from Web users, typically to commercial pages

questionnaire presentations of results of questionnaires

QA Q&A, FAQ, ...

list lists of words, numbers, etc

links lists of links to webpages with simple descriptions

top “top” pages that typically present the menu/structure of sites

speech speech or interview transcripts

errors pages that are not readable due to encoding problems, duplicates of other retrieved pages in the same corpus, pages with no contents

others cover class for genres represented by very few documents

Note that we broke down information and essay into subcategories depending on rhetorical types (i.e., argumentative, instructional etc.), being inspired in part by Santini (2005). We also distinguished journalistic from non-journalistic news, e.g., school or community news (news and njnews, respectively), and academic reports from non-academic ones (acreport and report, respectively). Finally, note the difference between info and report: the former pertains to information about a certain topic, e.g., information about some concert (the time and place of the event, etc.), while the latter presents contents that directly pertain to the topic, e.g., a report that presents the experience of going to the concert. We originally used more than the 27 classes reported above, but for ease of post-classification analysis, we collapsed categories with less than 3 pages in any corpus into the *others* category.

3.2 Results: Domains

3.2.1 Effects of time interval: Genki 2004 vs. Genki 2005

Since the Genki 2004 and 2005 corpora were constructed with the same procedure and with the same seed terms, but at different times (June 2004 and April 2005, respectively), the comparison of the two Genki corpora in terms of distribution of topic domains allows us to examine specifically how the time interval factor, which is 10 months in this case, affects the distribution of topic domains. The downloaded webpages were distributed across domains as shown in table 2, where the number and percentage of documents and their average size in number of tokens are summarized for each Genki corpus. The percentage values are also plotted in figure 1.

	Genki 2004			Genki 2005		
	# of docs	%	avg. size	# of docs	%	avg. size
appsci	24	2.7	2,451	8	2.7	3,914
arts	41	4.6	6,313	14	4.7	3,167
business	219	24.5	2,564	53	17.7	2,245
error	47	5.3	4,522	18	6	13,396
leisure	185	20.7	3,706	68	22.7	3,557
life	284	31.8	4,586	109	36.3	4,611
natsci	10	1.1	3,328	1	0.3	1,640
politics	7	0.8	5,826	1	0.3	1,573
socsci	77	8.6	4,151	28	9.3	8,564
total	894	100	3,885	300	100	4,744

Table 2. Distribution of topic domains in the Genki 2004 and 2005 corpora

Here we see that in both corpora life, business and leisure are the three major domain types, although there is a difference in ranking: life > business > leisure in 2004; life > leisure > business in 2005. This suggests an increase in the proportion of "personal interest" pages. The other domains are distributed in a more or less similar manner in the two corpora, as shown in figure 1. Some differences are found between the two Genki corpora, but we conclude that the effect of time interval is not very strong, since the two corpora share major characteristics, i.e., overall dominance of "personal interest" and commercial pages.

Comparing our results with the ones of Sharoff (for corpora in English, Russian, German), we notice that the total percentage of socsci and politics is only about 10% in our corpora, while his corpora overall show higher percentages, ranging from 15% to 29% in the three languages. Another difference is that our Genki corpora show a higher percentage of documents about life and leisure that refer to everyday life topics or personal interests. In our Genki corpora, the sum of life and leisure is consistently higher than 50% (52.5% in 2004, 59% in 2005), while in Sharoff's corpora the value ranges from 25% (English) to 51% (Russian). We suspect that these two differences between Sharoff's corpora and

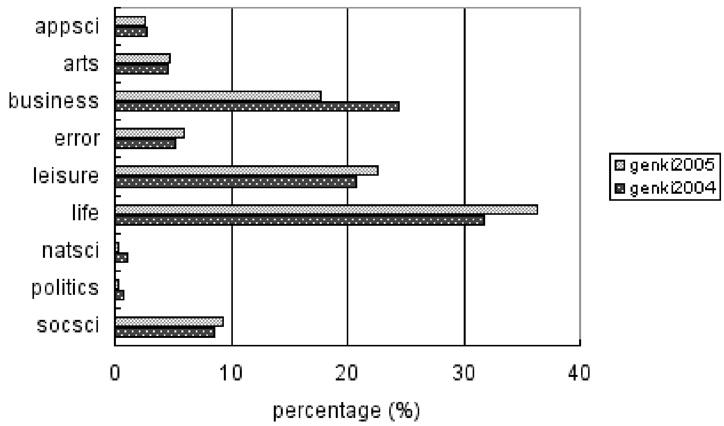

Figure 1. Percentage distribution of topic domains in the Genki 2004 and 2005 corpora

our corpora are mainly due to differences in seed choice. Our seeds, having been extracted from a basic vocabulary list, are more often related to everyday life domains, whereas Sharoff's seeds come from existing traditional corpora, and thus they tend to reflect some of the "higher" domains attested in these corpora. In the next section, we will investigate effects of seed selection by comparing the distribution of topic domain types in the Genki 2005 and BNC-seeded 2005 corpora.

3.2.2 Effects of seed selection: Genki 2005 vs. BNC-seeded 2005

The distribution of topic domain types is summarized in table 3, where the number and percentage of documents and their average size in number of tokens are presented for each domain type for the Genki 2005 and BNC-seeded 2005 corpora. The percentage values are also plotted in figure 2. Genki 2005 and BNC-seeded 2005 show more differences in domain distributions than the two

	Genki 2005			BNC-seeded 2005		
	# of docs	%	avg. size	# of docs	%	avg. size
appsci	8	2.7	3,914	17	5.7	3,702
arts	14	4.7	3,167	15	5	8,469
business	53	17.7	2,245	75	25	2,465
error	18	6	13,396	15	5	4,480
leisure	68	22.7	3,557	36	8.7	7,684
life	109	36.3	4,611	30	10	6,813
natsci	1	0.3	1,640	21	7	2,957
politics	1	0.3	1,573	65	21.7	6,037
socsci	28	9.3	8,564	36	12	7,103
total	300	100	4,744	300	100	5,188

Table 3. Distribution of topic domains in the Genki 2005 and BNC-seeded 2005 corpora

Genki corpora. With respect to Genki 2005, the proportions of five topic domains, appsci, business, natsci, politics, socsci – and the latter two in particular – are much higher than in BNC-seeded 2005. A decrease in leisure and life appears to be a trade-off of this increase. These differences cue two general changes that are likely to be caused by the change of seeds: an increase in the proportion of scientific and socio-political pages, and a decrease in the proportion of "personal interest" pages.

Strictly speaking, the comparison of Genki 2005 and BNC-seeded 2005 is not the best way of investigating effects of seed selection by excluding effects of time interval, since the two corpora were not constructed at the same time: the Genki 2005 corpus was built in April 2005, the BNC-seeded 2005 in August 2005. However, considering that the differences between the two corpora (at a 4-month interval) are much greater than those between the two Genki corpora (at a 10-month interval), we believe it is safe to conclude that the distribution of topic domain types in a Web corpus depends more on seed selection than on time interval.

Comparing our BNC-seeded 2005 corpus with Sharoff's En-

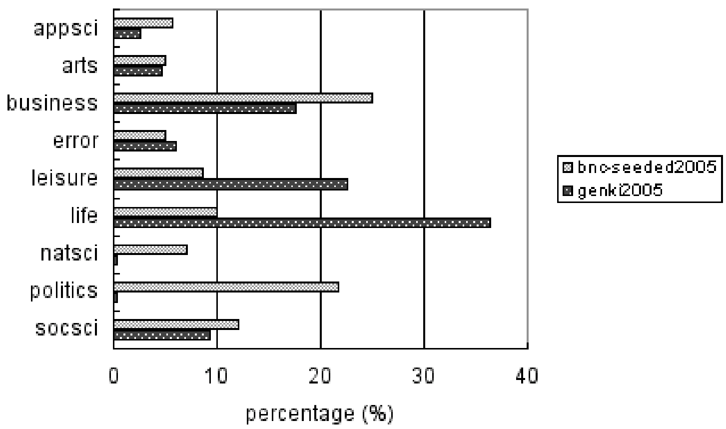

Figure 2. Percentage distribution of topic domains in the Genki 2005 and BNC-seeded 2005 corpora

glish Web corpus it is appropriate to examine similarities and differences between English and Japanese in the distribution of domain types. The reasoning is that the two corpora were built with more or less the same automated procedure and with similar seeds (we picked 100 words randomly from Sharoff's English word list), although, again, the corpora were not constructed at the same time, and, of course, there may also be effects due to the difference in annotators. The percentage distribution of domain types in our Japanese corpus (BNC-seeded 2005) and his English corpus (I-EN) is presented in figure 3. There are several notable differences. Two major domains in BNC-seeded 2005 are business and politics, as opposed to appsci and socsci in the I-EN corpus. Sharoff reported that in the I-EN corpus the majority of socsci pages are legal texts (legislation, law reports, terms and conditions, etc.), but we found almost no case of legal text in the BNC-seeded 2005, where a majority of pages labeled as socsci belong to other subdomains such as sociology, education or

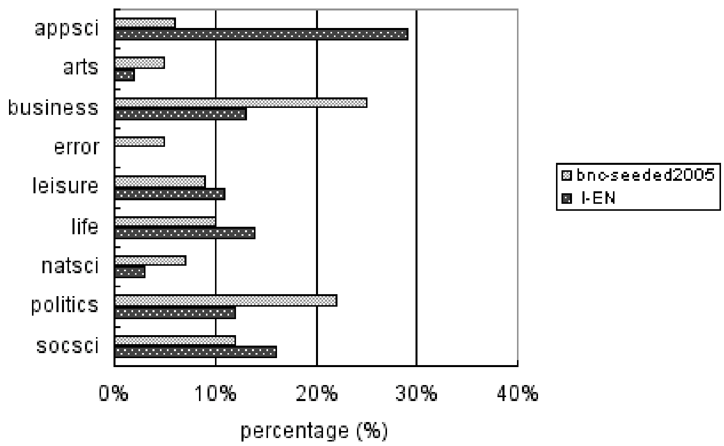

Figure 3. Percentage distribution of domain types in Sharoff's English corpus (I-EN) and our Japanese corpus (BNC-seeded 2005)

language. For the other domain types, we found no obvious difference. These results suggest that Web documents in different languages (at least, English and Japanese as indexed by Google) differ in the distribution of topic domains.

3.3 Results: Genre types

3.3.1 Effects of time interval: Genki 2004 vs. Genki 2005

The distribution of genre types in the two Genki corpora is presented in table 4, which summarizes the number and percentage of documents and their average size in number of tokens for each genre type. The percentage values are also plotted in figure 4. The general pattern that we found here is that in both corpora the genre types typical of personal prose – i.e., BBS, blog, diary, essay and personal – occupy a good portion of the distribution. The sum of these genres is 39.9% in Genki 2004 and 49% in Genki 2005. The overall dominance of the personal genres indicates that the Web-based corpora are likely to include a good amount of spontaneous prose produced by non-professional writers, which seems to match the dominance of “personal interest” pages in the results of the domain evaluation of the Genki corpora presented in section 3.2.1. Since this type of prose is not available in traditional corpora, Web-based corpora can be a very precious new linguistic resource.

Interestingly, we notice a sharp increase in the overall proportion of these genres between 2004 and 2005, suggesting the possibility that the Japanese Web (at least as ranked by Google and retrieved with our method) is becoming richer in personal prose. Another prominent genre type is commerinfo (commercial information). It occupies 18.6% and 14% of Web documents in the Genki 2004 and 2005 corpora, respectively (indicating that, at least according to our sample, its overall share is receding, perhaps in correspondence with the increase in personal pages). Together, personal and commercial pages constitute the majority of our Web-based corpora. The sum of these two types is 58.5% and 63% in 2004 and 2005, respectively. In contrast, the ratio of news is surprisingly low (1.1% in 2004, 0% in 2005), and there is no single case of acreport (reports of academic research) in either corpus. This may again be caused by our selection of seed terms, as was probably the case for the low percentage of politics and socsci in the results of the domain evaluation of the Genki corpora.

The genre types that tend not to include a good chunk of prose, such as links (links to other webpages), top (top pages with a site menu) and list (lists of words or numbers), have a relatively low ratio (8.6% in 2004 and 5.6% in 2005 in total). This is, of course, good news.

In summary, the genre evaluation of the Genki 2004 and 2005 corpora shows that a good majority of Web documents retrieved with Genki seeds are constituted by personal or commercial genres rather than academic or journalistic genres, which fits in nicely

	Genki 2004			Genki 2005		
	# of docs	%	avg. size	# of docs	%	avg. size
acreport	0	0	0	0	0	0
argessay	7	0.8	3,158	4	1.3	3,524
BBS	54	6.0	8,243	10	3.3	9,329
blog	55	6.2	3,959	74	24.7	4,604
comments	10	1.1	2,040	9	3.0	7,248
commerinfo	166	18.6	2,433	42	14.0	2,393
diary	165	18.5	5,019	47	15.7	5,284
error	51	5.7	4,171	18	6.0	13,396
essay	66	7.4	3,414	12	4.0	4,897
info	14	1.6	1,813	8	2.7	2,296
instinfo	32	3.6	2,790	9	3.0	3,588
links	48	5.4	1,768	7	2.3	2,327
list	15	1.7	4,949	6	2.0	550
magazine	13	1.5	4,332	0	0	0
news	10	1.1	3,316	0	0	0
njnews	5	0.6	5,109	3	1.0	1,426
novel	18	2.0	10,367	4	1.3	3,236
others	10	1.1	4,207	8	2.7	7,780
personal	16	1.8	2,138	4	1.3	1,909
QA	33	3.7	2,966	4	1.3	2,759
questionnaire	24	2.7	3,724	5	1.7	1,393
report	51	5.7	2,367	15	5.0	3,492
review	5	0.6	5,733	0	0	0
speech	5	0.6	9,131	4	1.3	2,671
teaching	8	1.9	5,362	3	1.0	3,741
top	13	1.5	1,623	4	1.3	2,893
total	894	100	3,885	300	100	4,744

Table 4. Distribution of genre types in the Genki 2004 and 2005 corpora

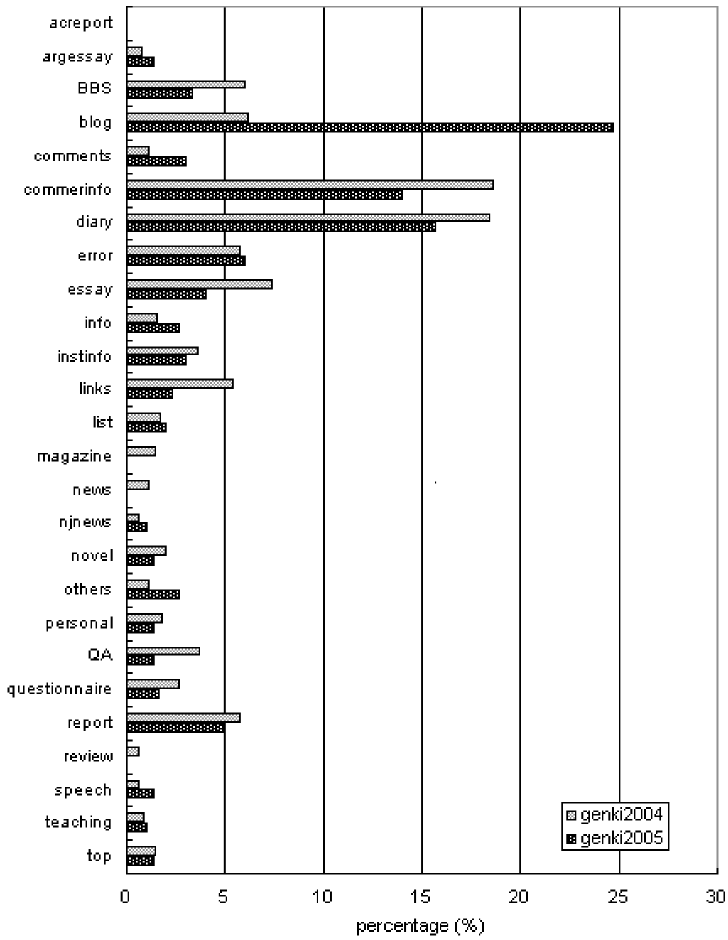

Figure 4. Percentage distribution of genre types in the Genki 2004 and 2005 corpora

with the results of the domain classification. This overall pattern is observed commonly in both corpora, although there are some differences, e.g., an increase in the proportion of personal genres, which suggests that the Japanese Web may be becoming richer in personal prose.

3.3.2 Effects of seed selection: Genki 2005 vs. BNC-seeded 2005

We also compared the Genki 2005 and BNC-seeded 2005 corpora in terms of the distribution of genre types in order to further examine effects of seed selection. The results of the genre evaluation are presented in table 5 and figure 5. Here we find some dramatic changes between the two corpora. In the BNC-seeded 2005, there is a sharp decrease in the proportion of pages of blog and diary, two major personal genres, while there is a substantial increase in the proportion of genres where academic, journalistic or public contents are presented (e.g., acreport, argessay, news and report). These changes in genre distribution match with the results of the domain evaluation that show an increase in the proportion of scientific and sociopolitical topics.

We notice that the magnitude of the changes between the Genki 2005 and BNC-seeded 2005 corpora in the genre type distribution is much greater than that between the two Genki corpora. Considering this finding, we believe that the distribution of genre types in the Web corpus largely depends on the nature of seed selection just like in the case of the distribution of domain types (see section 3.2).

3.4 Discussion

We have manually classified webpages of our three Japanese Web corpora in terms of domains and genres to examine how time interval and seed selection affect characteristics of the resulting Web

	Genki 2005			BNC-seeded 2005		
	# # of docs	%	avg. size	# of docs	%	avg. size
acreport	0	0	0	8	2.7	11,172
argessay	4	1.3	3,524	25	8.3	4,916
BBS	10	3.3	9,329	4	1.3	19,757
blog	74	24.7	4,604	19	6.3	7,228
comments	9	3.0	7,248	3	1.0	1,325
commerinfo	42	14.0	2,393	48	16.0	1,693
diary	47	15.7	5,284	16	5.3	8,079
error	18	6.0	13,396	15	5.0	4,480
essay	12	4.0	4,897	11	3.7	6,179
info	8	2.7	2,296	21	7.0	3,325
instinfo	9	3.0	3,588	10	3.3	3,324
links	7	2.3	2,327	0	0	0
list	6	2.0	550	5	5	7,876
magazine	0	0	0	12	4	8,039
news	0	0	0	13	4.3	6,065
njnews	3	1.0	1,426	4	1.3	5,418
novel	4	1.3	3,236	5	1.7	14,522
others	8	2.7	7,780	9	3.0	3,868
personal	4	1.3	1,909	18	6.0	6,517
QA	4	1.3	2,759	0	0	0
questionnaire	5	1.7	1,393	0	0	0
report	15	5.0	3,492	36	12.0	3,320
review	0	0	0	0	0	0
speech	4	1.3	2,671	6	2.7	4,248
teaching	3	1.0	3,741	4	2.0	348
top	4	1.3	2,893	8	1.3	11,172
total	300	100	4,744	300	100	5,188

Table 5. Distribution of genre types in the Genki 2005 and BNC-seeded 2005 corpora

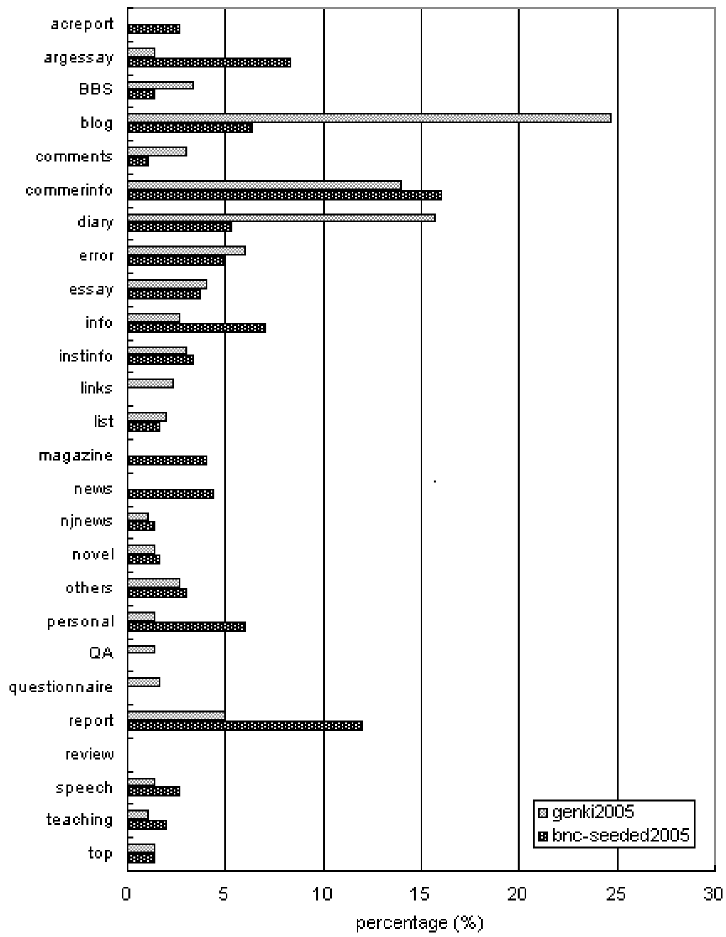

Figure 5. Percentage distribution of genre types in the Genki 2005 and BNC-seeded 2005 corpora

corpora. The two main findings have been as follows: 1) both factors affect characteristics of Web corpora considerably; 2) however, the effect of seed selection is notably stronger than that of time. In consideration of the results of corpus classification, one might wonder if the general increase in corpus size and average document size both from Genki 2004 to Genki 2005 and from Genki 2005 to BNC-seeded 2005, which was reported in section 2, are due to differences in the domains/genres that characterize the various corpora. We thoroughly examined the distributions of sizes within domains and genres for each pair (Genki 2004 vs. 2005, and Genki vs. BNC-seeded 2005), but we did not find any systematic correlation between the average text size and the distribution patterns of domains and genres. This indicates that the general increase of the average corpus size is not caused by changes in distribution of text types in a systematic way. One possible alternative explanation is that a good number of webpages increases in size over time as new contents are added. It will be interesting to examine this possibility by observing chronological changes in text size for the same webpages.

4 Typical lexical items

In this section, we examine how time interval and seed selection affect Japanese Web corpus construction from a lexical point of view. For this purpose, we conducted a qualitative analysis of typical lexical items in our three Japanese Web corpora. For two pairs of our three Japanese Web corpora (Genki 2004 vs. 2005 and Genki vs. BNC-seeded 2005), we compared the frequency of occurrence of each “word” (as tokenized by ChaSen) in each corpus with its frequency in the other corpus by computing the log-likelihood ratio association measure (Dunning 1993). We then evaluated the lists of words ranked by log-likelihood ratio, focusing in particular on the top 300 items in each list (Sharoff applies the same methodology; see his article for a discussion of the log-likelihood ratio measure).

In the top lists of the two Genki corpora, we did not find any systematic difference except for the following. The Genki 2004 list contains more lexical items related to business or finance (e.g., *tenpo* "store", *gokakunin* "confirmation") – 29 relevant items in the top 300 list – while there are only 3 items in the top 300 list of the Genki 2005. This may be explained by the higher proportion of pages classified as business in Genki 2004 than in Genki 2005, as reported earlier. In contrast, some dramatic difference has emerged from the comparison of the top 300 word lists of the Genki 2005 and BNC-seeded 2005 corpora. The BNC-seeded 2005 list contains a high proportion of terms used in socio-political text, i.e., 43% of the list (e.g., *seefu* "government", *kenpoo* "constitution"), while no instance of this sort is found in the Genki 2005 list. The difference must be due to the change in seed selection that has caused a major boost in the proportion of socio-political text.

In summary, the analysis of the data ranked by log-likelihood ratio for the Genki 2004 and 2005 corpora did not show any fundamental differences, while a strong difference emerged from the results of the comparison between the Genki 2005 and BNC-seeded 2005 corpora. This indicates that seed selection impacts on the lexical distribution of the resulting corpus more than time interval, as it does with the composition of domains and genres (the phenomena are obviously related).

5 Conclusion

The qualitative evaluation of the Japanese Web corpora built with automatic queries to Google coherently shows the following two patterns: 1) both seed selection and time interval affect the distribution of text and lexicons in the resulting Web corpus; 2) the effect of seed selection is much stronger than the effect of time interval. The difference between the two examined factors in magnitude of effects may be partly explained by the fact that the two factors affect Web-based corpus construction in different ways. Seed selection directly pertains to the way in which we sample documents from the Web. However, this is not the case for time interval. Time interval is rather relevant to changes in extrinsic factors such as indexing and ranking of Web documents by search engines, modifications of webpage contents, and so on. Such extrinsic factors largely characterize the dynamic nature of Web documents, but the changes due to time interval between corpus construction sessions affect the overall distributional properties of the resulting Web-based corpora, in terms of domain, genre and lexicon, much less than seed selection. To further study this point, we would like to observe chronological changes by repeatedly constructing Web-based corpora with a certain fixed time interval and the same procedure used to build Genki 2004 and 2005.

The prominent effect of seed selection on Web corpus construction suggests that a good understanding of the cause-and-effect relation between seeds and retrieved documents is an important step to gain some control over the characteristics of Web-based corpora, in particular, for the construction of general-purpose or reference corpora that are meant to represent a language as a whole. This boils down to a need to understand distributional properties of Web documents and then find a good method to randomly sample a set of documents that represent those properties with minimal bias toward certain domains, and seed selection is a very crucial part of the automatic sampling process. As far as we know, this line of research has not been widely pursued yet, except for the preliminary experiments by Ciaramita and Baroni (this volume). They propose and test an automated, quantitative, knowledge-poor method to evaluate the randomness of a Web corpus (with respect to a number of non-random/biased partitioning of the whole collection of Web documents). The results of their experiments indicate some effect of seed frequency on the randomness of the resulting corpus: i.e., medium frequency seeds might lead to a less biased corpus than either high frequency terms or terms selected from the whole frequency range. This line of research is crucial for finding an effective automated method to construct general-purpose balanced corpora from the Web. We are interested in further testing the effect of different seed sets picked on the basis of frequencies, and semantic/topical domains (e.g, arts, leisure, life, politics, etc.), to see how the properties of seed sets correlate with the distributional properties and quality of the resulting corpus.

References

Banno, E., Onno, Y., Sakane, Y. and Shinagawa, C. (1999). *Genki: An integrated course in elementary Japanese.* Tokyo: The Japan Times.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of LREC 2004*, 1313-1316.

Baroni, M. and Ueyama, M. (2004). Retrieving Japanese specialized terms and corpora from the World Wide Web. *Proceedings of KONVENS 2004.*

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61-74.

Fetterly, D., Manasse, M., Najork, M. and Wiener, J. (2004) A large-scale study of the evolution of Web pages. *Software: Practice & Experience* 34, 213-237.

Fletcher, B. (2004). Making the Web more useful as a source for linguistic corpora. In Connor, U. and Upton, T. (eds.) *Corpus linguistics in North America 2002*, Amsterdam: Rodopi.

Ghani, R., Jones, R.and Mladenić, D. (2003). Building minority language corpora by learning to generate Web search queries. *Knowledge and Information Systems* 7(1), 56-83.

Goto, H. (2003). Linguistic theories and linguistic resources: corpora and other data (Gengo riron to gengo shiryoo: coopasu to coopasu igai no deeta). *Nihongogaku (Japanese Language Studies)* 22, 6-15.

Kilgarriff A. and Grefenstette, G. (2003). Introduction to the special issue on the Web as Corpus. *Computational Linguistics* 29(3), 333-347.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., and Asahara, M. (2000). Morphological analysis system ChaSen version 2.2.1 manual. NIST Technical Report.

Santini, M. (2005). Genres in formation? An exploratory study of Web pages using cluster analysis. Proceedings of CLUK 05.

Ueyama, M. and Baroni, M. (2005). Automated construction and evaluation of a Japanese Web-based reference corpus. *Proceedings of Corpus Linguistics 2005*, available online at `http://www.corpus.bham.ac.uk/PCLC/`.