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

Motoko Ueyama

SSLMIT, University of Bologna
136 Corso della Repubblica
47100 Forl, Italy
motoko@sslmit.unibo.it

Abstract

Recently, there have been efforts to construct written corpora by using the WWW. A promising approach to build Web corpora is to run automated queries to search engines and download pages found in this way. This makes it possible to build corpora rapidly and economically, but we cannot control what are contained in resulting corpora. Under these circumstances, it is important to verify the general nature of Web corpora. This study, in particular, investigated effects of two essential factors on three Japanese corpora that we built: seed terms used for queries; and time interval between different corpus construction sessions, which measures the stability of query results over time. We evaluated the corpora qualitatively, in terms of domains, genres and typical lexical items. Results show these two patterns: 1) both seed selection and time interval affect the distribution of text and lexicon; 2) the effect of seed selection is much stronger. The prominent effect of seed selection 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 corpora, in particular, for the construction of general corpora meant to represent a language as a whole.

1. Introduction

1.1. Automated methods to construct Web corpora

A considerable amount of work has been conducted on the use of the Web as a linguistic data source for various tasks (e.g., Kilgarriff and Grefenstette, 2003). A promising approach is to build corpora by running automated queries to search engines and post-processing the retrieved pages. This approach differs from the traditional method of corpus construction, where one needs to spend considerable time collecting the texts to include, but having good control on corpus contents. In contrast, the automated approach, despite the absence of quality control, makes it possible to construct corpora quickly and economically. This is good news for researchers in urgent need of large-scale balanced corpora for the language of their interest, but with no access to such corpora. This is the case for researchers working on the majority of the world's languages, including Japanese. The pioneering work in the automatic construction of Web corpora has been done by the CorpusBuilder project (e.g., Ghani et al., 2003). Baroni and Bernardini (2004) introduced the BootCaT tools, a free suite of Perl scripts for the automated construction of corpora via Google queries. The BootCaT tools were used for developing general corpora by Baroni and Ueyama (2004), Ueyama and Baroni (2005) and Sharoff (2006).

Despite a considerable amount of work on ways to use the Web as a linguistic data source, only few studies have evaluated Web corpora qualitatively. Fletcher (2004) found his English Web corpus to be characterized by a more interactive style, and to be more varied in comparison with the British National Corpus (henceforth BNC; Aston and Burnard, 1998). Sharoff (2006) adapted the BootCaT tools to build Web corpora including English, finding his English Web corpus to be richer in topic domains than the BNC. The findings of both studies challenge the view that Web corpora are not fit for linguistic research. In our last study (Ueyama and Baroni, 2005), we qualitatively evaluated two Japanese Web corpora built with the the BootCaT tools. On the basis of our manual classification data, we found in both corpora many documents to be produced by non-professional writers and characterized by everyday life topics and high occurrences of an informal, spontaneous, interactive style. This confirms the findings of Fletcher and Sharoff. We also found that this type of text was more dominant in our Japanese corpora than in any of Sharoff's corpora. This difference may be due to differences in seed choice. Many of our seeds pertain to everyday life, while this is not the case of Sharoff's seeds selected from existing traditional corpora (e.g. BNC).

1.2. Goals of the study

The difference between Sharoff's and our results leads us to the first goal of the study: investigate how different seed selection strategies affect the nature of resulting Web corpora. We conduct a qualitative investigation, analyzing two Japanese Web corpora that we built with seeds extracted from a basic Japanese vocabulary list and also from Sharoff's BNC-based English word list translated into Japanese. Another essential factor that possibly affects Web corpus construction is time interval. Search engine indexing continuously changes, which is expected to strongly affect query results, and, consequently, resulting Web corpora. Our second goal is to examine the effect of time interval in attempt to tackle the 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 same procedure and seeds. To investigate the effects of these two factors, 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 corpora and describe the characteristics of each corpus. In section 3., we describe our corpus clas-

sification methods and present our results, while section 4. presents the evaluation of typical lexical items of each corpus. Finally, in section 5. we discuss our findings and conclude by suggesting directions for future research.

2. Corpus construction

We built the Genki 2004 and 2005 corpora with the same procedure and seeds, but at two different times: July 2004 and April 2005, respectively (also analyzed in Ueyama and Baroni, 2005). The BNC-web 2005 corpus was built in August 2005, using the same procedure but different seeds.

2.1. Automated url search

For the two Genki corpora, to look for pages reasonably varied and not excessively technical, we decided to query a search engine (Google in our case) for words belonging to a basic Japanese vocabulary. Thus, we randomly picked 100 words from the word list of a Japanese Textbook, Genki (Banno et al., 1999). For the BNC-web 2005 corpus, we randomly selected 100 from Sharoff's (2006) list of 500 seeds extracted from the BNC (available at `http://corpus.leeds.ac.uk/internet/seeds-en`), and translated those in Japanese. Reflecting the nature of the BNC, seeds for the BNC-web 2005 corpus vary more in topics than the ones for the two Genki corpora.

All the three corpora were built with the BootCaT tools (Baroni and Bernardini, 2004). We randomly combined the 100 seeds into 100 triplets, and 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 with connected text (that contain at least 3 unrelated content words). We used the same triplets to build the two Genki corpora, while we created and used a new set of 100 triplets in for the BNC-web 2005 corpus. For each query, we retrieved maximally 10 urls from Google, and discarded duplicate urls. This gave us a total of 894 unique urls, 993, and 908 for Genki 2004, Genki 2005 and BNC-web 2005, respectively. Notice that, while for the purposes of this study we are satisfied with these sizes, the same procedure could be used to build larger corpora.

To find the number of common urls, we compared the two Genki corpora and found 187 urls in common, leaving 707 and 806 urls in the Genki 2004 only and in the Genki 2005 only, respectively. With respect to the Genki 2005 url list, the overlap is of less than 20%. To see if those urls were identical in contents, we randomly picked 20 out of the 187 common urls and examined the text, finding only 13 of the 20 urls (65%) to be identical. The overlap decreases even more between Genki 2005 and BNC-web 2005: only 11 urls (1% overlap with respect to the Genki 2005 url list).

2.2. Retrieving and post-processing text

The web page correponding to each url was automatically retrieved and formatted as text by removing "boilerplate" such as the HTML tags (using Perl's HTML::TreeBuilder as_text function and simple regular expressions). Since Japanese pages can be encoded in different character sets (e.g., shift-jis, euc-jp, utf-8), our script detects the character set of the page, using the HTML code, and converts that into utf-8. The ChaSen tool (Matsumoto at al, 2000) was used to tokenize the text. ChaSen expects input and outpu to be coded in euc-jp, while our text-processing scripts are designed for utf-8 input. We converted text back and forth between utf-8 and euc-jp with the recode tool (http://recode.progiciels-bpi.ca/). The tokenization results are shown in table 1. Comparing the two Genki corpora, we noticed that in Genki 2005 not only more and different urls but also urls with more text were retrieved, as indicated by the increased average document size. BNC-web 2005, 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 Genki 2005, although the total document count decreases. Finally, we found that some pages did not contain any substantial text: e.g., the ones not decoded properly. The ratio of such pages was about 5% for all the corpora. We believe that this error rate is tolerable in the sense that the great majority of text is usable.

	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-web 2005	908	5,732,080	6,313	5%

Table 1: Total documents, total tokens, average size, and error rate of the Genki 2004, Genki 2005, and BNC-web 2005 corpora

3. Corpus classification

We manually classified all 894 pages of Genki 2004, and 300 randomly picked from each of Genki and BNC-web 2005, in terms of domains and genres. For the domain analysis, we adopted Sharoff's (2006) BNC-based classification system, so that we can compare our results directly with his. We used the nine categories listed in table 2. For the genre analysis, we first went through a good amount of the Web pages to have a general idea about genre distributions, and selected 27 types (see table 3). We split each of info and essay into sub-categories depending on rhetorical types (i.e., argumentative, instructional...). We also distinguished journalistic from non-journalistic news (news and njnews), and academic from non-academic reports (acreport and report). Note the difference between info and report: the former pertains to information about a certain topic, e.g., concert information (the time and place of the event, etc.), while the latter presents contents relevant to the topic, e.g., a report about the experience of going to the concert.

3.1. Results: domains

3.1.1. Effects of time interval

Since we built the two Genki corpora with the same procedure and seeds, but at different times, the comparison of this pair allows us to examine how time inverval affects domain distributions. The classification results are summarized in table 4, where the number and percentage of documents and their average size in token number are shown for each

natscsi	agriculture, astronomy, meteorology, ...
appsci	computing, engineering, medicine, transport, ...
socsci	law, history, sociology, language, education, ...
politics	
business	e-commerce pages, company homepages, ...
life	general topics related to everyday life
arts	literature, visual arts, performing arts, ...
leisure	sports, travel, entertainment, fashion, hobbies ...
error	encoding errors, duplicates, ...

Table 2: Domain types

blog	personal pages created by users registered at blog servers
BBS	bulletin boards with discussion pages
diary	an "adaptive" genre also existing in traditional texts (see Santini, 2005)
personal	personal pages not created through a blog service
argessay	essays in an argumentative rhetoric style
essay	essays in a non-argumentative style
novel	another adapted genre type
commerinfo	pages to promote services/sell products
instrinfo	pages to help to perform a certain task
info	non-commercial info. pages (see below)
teaching	materials for instruction
news	traditional journalistic news
njnews	non-journalistic news, such as community pages, Web magazine
acreport	reports of academic research
report	with contents related to a certain topic (see below)
review	product/service evaluation, critique of of artistic activies
comments	sent from Web users to commercial pages
questionnaire	presentations of results of questionnaires
QA	Q&A, FAQ, ...
list	lists of words, numbers, etc
links	links to Web pages with simple descriptions
top	"top" pages that present the menu of sites
speech	speech or interview transcripts
errors	pages with no sublstantial text/contents
others	class for genres with very few documents

Table 3: Genre types

corpus. The percentage values are also plotted in figure 1. In both corpora, life, business and leisure are the three major domains, and in Genki 2005 there is an increase in life and leisure that both contain " personal intrest" pages. The other domains are distributed in a similar way. Despite some differences, we conclude that the effect of time interval is not very strong, since the two corpora share a major characeristic, i.e., the dominance of "personal interest" and commercial pages.

In comparison with Sharoff's (2006) results for corpora in English, Russian, German, the total percentage of socsci and politics is only about 10% in our Genki corpora, while his corpora show higher percentages, ranging from 15% to 29% in the three languages. Another difference is that in our corpora the sum of life and leisure is higher than 50%, while in Sharoff's corpora the value ranges from 25% (English) to 51% (Russian). These differences are likely to be

	Genki 2004		Genki 2005	
	# of docs	avg. size	# of docs	avg. size
appsci	24 (2.7%)	2451	8 (2.7%)	3914
arts	41 (4.6%)	6313	14 (4.7%)	3167
business	219 (24.5%)	2564	53 (17.7%)	2245
error	47 (5.3%)	4522	18 (6%)	13396
leisure	185 (20.7%)	3706	68 (22.7%)	3557
life	284 (31.8%)	4586	109 (36.3%)	4611
natsci	10 (1.1%)	3328	1 (0.3%)	1640
politics	7 (0.8%)	5826	1 (0.3%)	1573
socsci	77 (8.6%)	4151	28 (9.3%)	8564
total	894 (100%)	3885	300 (100%)	4744

Table 4: Distribution of topic domains in the Genki 2004 and 2005 corpora

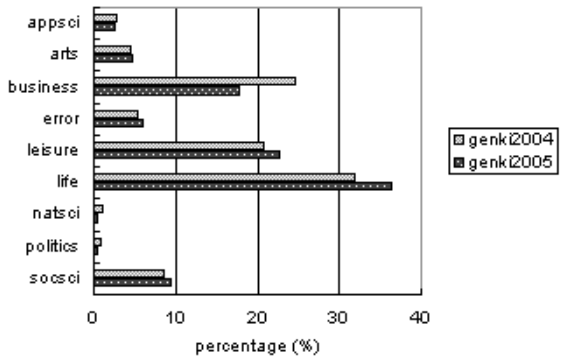

Figure 1: Percentage distribution of topic domains in the Genki 2004 and 2005 corpora.

mainly caused by seed differences. Most of our seeds are related to everyday life, whereas Sharoff's seeds tend to reflect some of the "higher" domains attested in his corpora.

3.1.2. Effects of seed selection

To investigate seed selection effects on domain distributions, we compared Genki and BNC-web 2005. Results are summarized in table 5, and the percentage values are plotted in figure 2. Genki and BNC-web 2005 show more dramatic differences than the two Genki corpora. BNC-web 2005 shows higher proportions of appsci, business, natsci, politics, and socsci, and a decrease in leisure and life. These cue two changes that are likely to be caused by the change of seeds: an increase in scientific and socio-political pages, and a decrease in "personal interest" pages. Strictly speaking, the comparison of these two corpora is not ideal to find seed selection effects by excluding time interval effects, since they were not built at the same time. However, considering the greater differences between these two (at a 4-month interval) than those between the two Genki corpora (at a 10-month interval), it seems to be OK to conclude that the domain distribution depends more on seed selection than on time interval.

Comparing BNC-web 2005 with Sharoff's (2006) English Web corpus is appropriate to examine differences between English and Japanese in domain distributions. The two corpora were built with more or less the same automated procedure and with the similar seeds (our seeds were picked

	Genki 2005		BNC-web 2005	
	# of docs	avg. size	# of docs	avg. size
appsci	8 (2.7%)	3914	17 (5.7%)	3702
arts	14 (4.7%)	3167	15 (5%)	8469
business	53 (17.7%)	2245	75 (25%)	2465
error	18 (6%)	13396	15 (5%)	4480
leisure	68 (22.7%)	3557	36 (8.7%)	7684
life	109 (36.3%)	4611	30 (10%)	6813
natsci	1 (0.3%)	1640	21 (7%)	2957
politics	1 (0.3%)	1573	65 (21.7%)	6037
socsci	28 (9.3%)	8564	36 (12%)	7103
total	300 (100%)	4744	300 (100%)	5188

Table 5: Distribution of topic domains in the Genki 2005 and BNC-web 2005 corpora

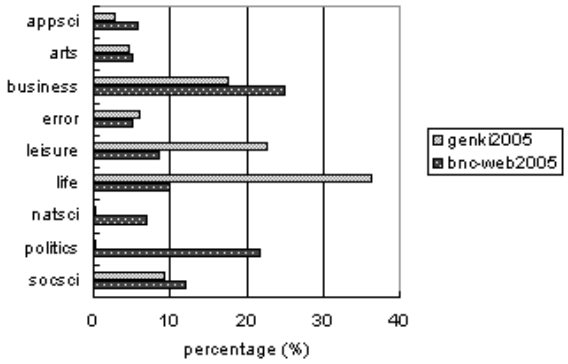

Figure 2: Percentage distribution of topic domains in the Genki 2005 and BNC-web 2005 corpora.

from his seed list), although not constructed at the same time. The domain distribution is plotted for our BNC-web 2005 corpus and his English corpus (I-EN) in figure 3. We find two main differences. First, two major domains in BNC-web 2005 are business and politics, while appsci and socsci in I-EN. Second, in I-EN, most pages of socsci are legal texts, but there is almost no case of this text type in BNC-web 2005, where the majority of sosci pages belong to other subdomains, e.g., sociology or education. For the other domains, we found no obvious difference. These results suggest that Web documents in different languages can vary in domain distirubtions.

3.2. Results: genres

3.2.1. Effects of time interval

The genre distributions in the two Genki corpora are presented in table 6, and the percentage values are plotted in figure 4. In both corpora, a good portion of the distribution is occupied by the genres typical of personal prose, i.e., BBS, blog, diary, essay and personal pages. The sum of these genres is 39.9% in Genki 2004 and 49% in Genki 2005. This indicates that the web corpora are likely to include a good amount of spontaneous prose composed by non-professional writers. This pattern matches with the dominance of "personal interest" pages in the results of the domain analysis of the Genki corpora. Since this text type is not available in traditional corpora, web corpora can be a very precious new linguistic resource. We also notice a

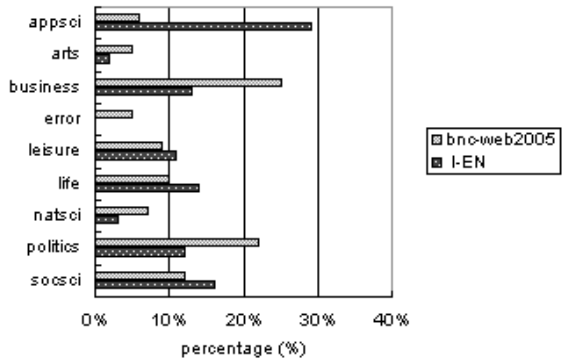

Figure 3: Percentage distribution of domain types in our Japanese corpus (BNC2005) and Sharoff's English corpus (I-EN).

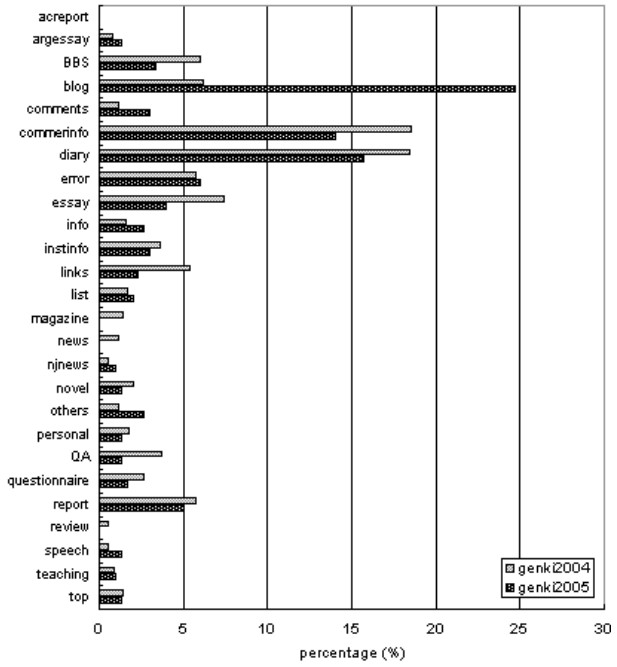

Figure 4: Percentage distribution of genre types in the Genki 2004 and 2005 corpora.

sharp increase in the overall ratio of these genres between 2004 and 2005, suggesting the possibility that the Japanese Web is becoming richer in personal prose. Another prominent genre is commerinfo: 18.6% in 2004 and 14% of documents in 2005. The sum of all these genres is 58.5% in 2004 and 63% in 2005, which indicates the dominacne of personal and commercial genres. In contrast, the ratio of news is very low, and there is no single case of acreport (academic research reports) in either corpus. This may again be caused by our seed selection, as was probably the case for the low proportion of politics and socsci in the results of the domain analysis of the Genki corpora. The genres with almost no good chunk of prose, such as links, top and list have a relatively low ratio, which is good news.

3.2.2. Effects of seed selection

We also compared Genki and BNC-web 2005 in terms of genre distributions to further examine seed selection ef-

	Genki 2004		Genki 2005	
	# of docs	avg. size	# of docs	avg. size
acreport	0 (0%)	0	0 (0%)	0
argessay	7 (0.8%)	3158	4 (1.3%)	3524
BBS	54 (6.0%)	8243	10 (3.3%)	9329
blog	55 (6.2%)	3959	74 (24.7%)	4604
comments	10 (1.1%)	2040	9 (3.0%)	7248
commerinfo	166 (18.6%)	2433	42 (14.0%)	2393
diary	165 (18.5%)	5019	47 (15.7%)	5284
error	51 (5.7%)	4171	18 (6.0%)	13396
essay	66 (7.4%)	3414	12 (4.0%)	4897
info	14 (1.6%)	1813	8 (2.7%)	2296
instinfo	32 (3.6%)	2790	9 (3.0%)	3588
links	48 (5.4%)	1768	7 (2.3%)	2327
list	15 (1.7%)	4949	6 (2.0%)	550
magazine	13 (1.5%)	4332	0 (0%)	0
news	10 (1.1%)	3316	0 (0%)	0
njnews	5 (0.6%)	5109	3 (1.0%)	1426
novel	18 (2.0%)	10367	4 (1.3%)	3236
others	10 (1.1%)	4207	8 (2.7%)	7780
personal	16 (1.8%)	2138	4 (1.3%)	1909
QA	33 (3.7%)	2966	4 (1.3%)	2759
questionnaire	24 (2.7%)	3724	5 (1.7%)	1393
report	51 (5.7%)	2367	15 (5.0%)	3492
review	5 (0.6%)	5733	0 (0%)	0
speech	5 (0.6%)	9131	4 (1.3%)	2671
teaching	8 (1.9%)	5362	3 (1.0%)	3741
top	13 (1.5%)	1623	4 (1.3%)	2893
total	894 (100%)	3885	300 (100%)	4744

Table 6: Distribution of genre types in the Genki 2004 and 2005 corpora

	Genki 2005		BNC-web 2005	
	# of docs	avg. size	# of docs	avg. size
acreport	0 (0%)	0	8 (2.7%)	11172
argessay	4 (1.3%)	3524	25 (8.3%)	4916
BBS	10 (3.3%)	9329	4 (1.3%)	19757
blog	74 (24.7%)	4604	19 (6.3%)	7228
comments	9 (3.0%)	7248	3 (1.0%)	1325
commerinfo	42 (14.0%)	2393	48 (16.0%)	1693
diary	47 (15.7%)	5284	16 (5.3%)	8079
error	18 (6.0%)	13396	15 (5.0%)	4480
essay	12 (4.0%)	4897	11 (3.7%)	6179
info	8 (2.7%)	2296	21 (7.0%)	3325
instinfo	9 (3.0%)	3588	10 (3.3%)	3324
links	7 (2.3%)	2327	0 (0%)	0
list	6 (2.0%)	550	5 (5%)	7876
magazine	0 (0%)	0	12 (4%)	8039
news	0 (0%)	0	13 (4.3%)	6065
njnews	3 (1.0%)	1426	4 (1.3%)	5418
novel	4 (1.3%)	3236	5 (1.7%)	14522
others	8 (2.7%)	7780	9 (3.0%)	3868
personal	4 (1.3%)	1909	18 (6.0%)	6517
QA	4 (1.3%)	2759	0 (0%)	0
questionnaire	5 (1.7%)	1393	0 (0%)	0
report	15 (5.0%)	3492	36 (12.0%)	3320
review	0 (0%)	0	0 (0%)	0
speech	4 (1.3%)	2671	6 (2.7%)	4248
teaching	3 (1.0%)	3741	4 (2.0%)	348
top	4 (1.3%)	2893	8 (1.3%)	11172
total	300 (100%)	4744	300 (100%)	5188

Table 7: Distribution of genre types in the Genki 2005 and BNC-web 2005 corpora

fects. The results are presented in table 7 and figure 5. Some dramatic changes emerge from the results. BNC-web 2005 shows a sharp decrease in blog and diary, while there is a substantial increase in genres where academic, journalistic or public contents are presented, e.g., acreport, argessay, news and report. These changes match with the results of the domain analysis of the same corpora showing an increase in scientific and sociopolitical topics in BNC-web 2005. We also notice that the magnitude of the changes between Genki and BNC-web 2005 in the genre distribution is much greater than that between the two Genki corpora, in correspond with the results of the domain classification.

4. Typical lexical items

To examine how time interval and seed selection affect Web corpora lexically, we examined typical lexical items in our three Japanese corpora. For the two pairs (Genki 2004 vs. 2005 and Genki vs. BNC-web 2005), we compared the frequency of occurrence of each "word" (as tokenized by ChaSen) by computing the log-likelihood ratio association measure (Dunning, 1993; Sharoff, 2006), and evaluated the lists of words ranked by that measure, focusing on the top 300 items in each list. In the lists of both Genki corpora, we did not find any systematic difference except for the following one. The Genki 2004 list contains more items related to business or finance (e.g., *tenpo* "store," *gokakunin* "confirmation") – 29 instances out of the top 300 list – while

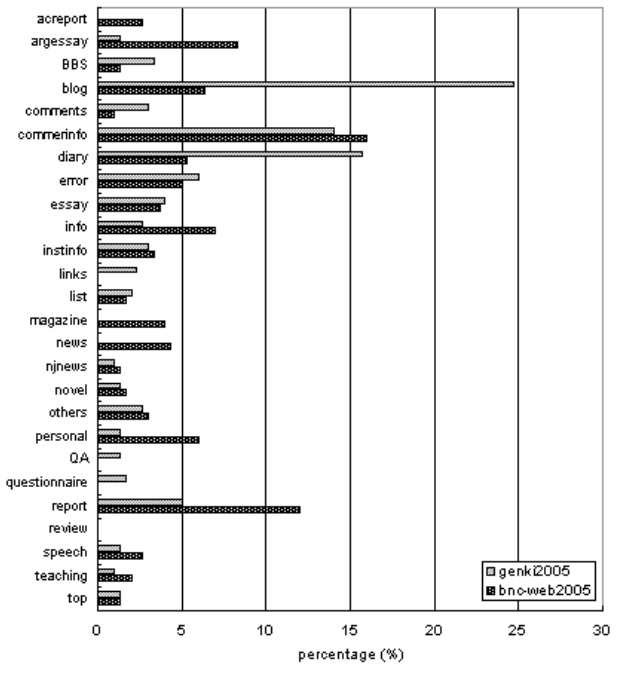

Figure 5: Percentage distribution of genre types in the Genki 2005 and BNC-web 2005 corpora.

we find only 3 in the Genki 2005 list. This may be explained by the higher ratio of business pages in Genki 2004, as reported earlier. In contrast, some dramatic difference has emerged from the comparison of Genki and BNC-web 2005. The BNC-web 2005 list contains a high ratio 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. This must be due to the change of seed selection that has caused the increase of socio-political text in BNC-web 2005. These results indicate that seed selection also impacts on the lexical distribution much more than time interval.

5. Discussion and conclusion

The qualitative evaluation of our Japanese Web corpora coherently shows the following: 1) both seed selection and time interval affect the nature of the resulting Web corpus; 2) the effect of seed selection is much stronger than that of time interval. The difference between the two examined factors may be partly explained in the following way. Seed selection directly pertains to the way of sampling documents from the Web, but this is not the case for time interval. Time interval rather relates to changes in extrinsic factors of Web documents, such as indexing and ranking of Web documents by search engines, web-page updates, and so on. Such extrinsic factors largely characterize the dynamic nature of Web documents, but they impact on the overall distribution of the resulting Web corpora much less than seed selection. It may be interesting to observe chronological changes by repeatedly constructing Web corpora with a certain fixed time interval and the same procedure used to build Genki 2004 and 2005.

The prominent effect of seed selection 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 nature of Web corpora, particularly, for constructing general balanced corpora 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 crucial part of this process. As far as we know, this line of research has not been widely pursued yet, except for the preliminary study by Ciaramita and Baroni (2006). They propose and test an automated, quantitative, knowledge-poor method to evaluate the randomness of a Web corpus, and their results indicate that medium frequency seeds might lead to a less biased corpus than either high frequency terms or terms selected from the whole frequency range. We are interested in further testing the effect of different seed sets picked on the basis of frequencies and topic domains, to see how the properties of seed sets correlate with the distributional properties of the resulting corpus.

6. References

G. Aston, and L. Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

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

M. Baroni and S. Bernardini. 2004. BootCaT: Boot strapping corpora and terms from the Web. *Proceedings of the Fourth Language Resources and Evaluation Conference*, Lisbon, Portugal.

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

M. Ciaramita and M. Baroni. 2006. A figure of merit for the evaluation of Web-corpus randomness. *Proceedings of EACL 2006*.

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

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

W. Fletcher. 2004. Making the Web more useful as a source for linguistic corpora. In U. Cornnor and T. Upton, editors, *Corpus Linguistics in North America 2002: Selections from the Fourth North America Symposium of the American Association for Applied Corpus Linguistics (Amsterdam: Rodopi)*. Available on-line from `http://miniappolis.com/KWiCFinder/`

R. Ghani, R. Jones, and D. Mladenic. 2003. *Building minority language corpora by learning to generate Web search queries, Knowledge and Information Systems 2003*.

H. Goto. 2003. Linguistic theories & linguistic resources: corpora and other data (Gengo riron to gengo shiryoo: coopasu to coopasu igai no deeta). *Nihongogaku (Japanese Language Studies)*, 22: 6-15. Available on-line from `http://www.sal.tohoku.ac.jp/~gothit/nhnggk0304.html`

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

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

M. Santini. 2005. Genres in formation? An exploratory study of Web pages using cluster analysis. Proceedings of CLUK 05. Available from `http://www.itri.brighton.ac.uk/~Marina.Santini/`

S. Sharoff. 2006. Creating general-purpose corpora using automated search engine queries.

M. Ueyama and M. Baroni. 2005. Automated construction and evaluation of a Japanese web-based reference corpus. *Proceedings of Corpus Linguistics 2005, Birmingham, U.K.*