Automated Construction and Evaluation of Japanese Web-based Reference Corpora

Motoko Ueyama & Marco Baroni
SSLMIT
University of Bologna
{motoko, baroni}@sslmit.unibo.it

1. Introduction

The World Wide Web, being essentially an enormous database of mostly textual documents, offers great opportunities to corpus linguists. An increasing number of studies has shown how Web-derived data are useful in a variety of linguistic and language technology tasks (Kilgarriff and Grefenstette 2003).

A particularly promising approach to the use of the Web for linguistic research is to build corpora via automated queries to search engines, retrieving and post-processing the pages found in this way (Ghani et al. 2003, Baroni and Bernardini 2004, Sharoff to appear). 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 has perfect control and awareness over contents. With automated Web-based corpus construction, the situation is reversed: one can build a corpus in very little time, but without a good control over what kinds of texts are in the corpus.

The automated methods, despite the almost complete absence of quality control, has made it possible to construct corpora for linguistic research in a quick and economic manner. This is good news for researchers who have no access to large-scale balanced corpora (i.e., something equivalent to the BNC) for the language of their interest, as is the case for most researchers working on almost all languages, including Japanese (see Goto 2003 for a survey of current availability of Japanese corpora for research purposes).

In this paper, we describe two Japanese “reference” corpora (“reference” in the sense that they are not meant to represent a specialized language, but Japanese in general, or at least the kind of Japanese written on the Web) that we have constructed using the aforementioned automated methods. The corpora contain about 3.5 million tokens and about 4.5 million tokens, respectively. The main goal of the current paper is to provide a relatively in-depth evaluation of the contents of the corpora, by manual classification of all the pages in the first corpus and of a sample of pages from the second one. The results of our evaluation indicate that these Web corpora are characterized by an abundance of relatively spontaneous, often interactive prose dealing with everyday life topics.

Moreover, since the two corpora were built using the very same methods at 10 months' distance from one another (in July 2004 and April 2005, respectively), we can also present some data related to the important issue of how “stable” the results of search engine queries are over time. We discovered that there is very little overlap between the pages retrieved in the two rounds, and that there is also some interesting development in terms of the typology of pages that were found. On the one hand, this suggests that the methodology can be promisingly applied to short time diachronic studies. On the other hand, it also indicates that different linguistic studies based on Web data, if they are meant to be comparable, should use the same set of retrieved pages (the same static corpus), rather than dynamically obtained data collected at different times.

The rest of the paper is structured as follows: In Section 2, we shortly review some related studies. In Section 3, we present the corpus construction procedure. In Section 4, we describe the domain categories and the genre type categories that we used to evaluate the Web pages of our corpora manually. In Section 5, we present the results of the evaluation of our Web corpora for the domain types and the genre types, and a more detailed analysis of the differences between the genres *diary* and *blog*. Finally, in section 6 we conclude by suggesting directions for further study.

2. Related work

There is by now a considerable amount of work on using the Web as a source of linguistic data (see, e.g., the papers collected in Kilgarriff and Grefenstette 2003). Here, we shortly review other studies that, like ours, used automated search engine queries to build a corpus.

The pioneering work in this area 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 less 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 classification of the pages.

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 (BNC, Aston and Burnard 1998) 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 5000 words in the BNC were lacking in 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.

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. They were used in this way by Baroni and Ueyama (2004), whose “reference corpus” is what we here call the “2004 corpus,” and Sharoff (to appear).

The work most closely related to this study is presented by Sharoff (to appear). He uses an adapted version of the BootCaT tools to build Web-derived corpora of more than 100M words for English, Russian and German. The corpora are constructed via automated Google queries for random combinations of 4-tuples of frequent words extracted from existing corpora. Sharoff classifies 200 randomly selected documents from each corpus in terms of various characteristics, including the domain of the document. He uses 8 domain categories inspired by the BNC classification (with some adaptations). In a comparison with the distribution of domains in the BNC, he finds that the English Web corpus (not surprisingly) is richer in technical, applied science domains, and poorer in texts from the arts (unfortunately, we do not have a balanced Japanese corpus available to use for similar comparisons). In our classification by domain, we adopted Sharoff's categories, so that our results are directly comparable. Sharoff also presents a comparison in terms of word frequencies between his Web corpora, reference corpora in English and Russian, and newswire corpora in English, Russian and German. He finds that the Web corpora are closer to the reference corpora than to the newswire corpora. His results also confirm Fletcher's findings about the Web being characterized by a more interactive style and more lexical variety.

While this paper has the relatively narrow goal to evaluate corpora automatically constructed with search engine queries in terms of what they have to offer for linguistic research, there is also a rich, relevant literature on the more general theme of identifying and defining “Web genres”, often with the goal to build genre-aware search engines (e.g., Meyer zu Eissen and Stein 2004 and Santini 2005).

It is also worth mentioning that, while as far as we know we are the first to present empirical work on how changes in search engine indexing across time can affect Web-based corpus construction, there is, of course, much interesting work on the evolution of the Web presented by the WWW/information retrieval community (e.g., Fetterly et al. 2004).

3. Corpus construction

Web built two Japanese corpora with the same automated procedure we are about to describe. One was constructed in July 2004, the other in April 2005.

In order to look for pages that were reasonably varied and not excessively technical, we considered that we should query Google for words belonging to the basic Japanese vocabulary. Thus, we randomly picked 100 words from the lexicon list of an elementary Japanese Textbook (Banno et al. 1999): e.g., **天気** *tenki* “weather,” **朝ご飯** *asagohan* “breakfast,” **スーパー** suupaa “supermarket,” **冷たい** *tsumetai* “cold”.

We then randomly combined these words 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 unrelated words from a basic vocabulary list). We used the very same triplets both in the July 2004 and in April 2005 for corpus construction.

For each query, we retrieved maximally 10 urls from Google, and we discarded duplicate urls. This gave us a total of 894 unique urls in 2004 and 993 in 2005.

Interestingly, only 187 urls were found in both rounds, leaving 707 urls that were retrieved in 2004 only and 806 urls that were retrieved in 2005 only. Thus, with respect to the 2005 url list, the overlap with the previous year is of less than 20%. Moreover, there is, of course no guarantee that the Web pages 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 the first author compared the 2004 and 2005 texts. We found that the two versions were identical in terms of contents only for 13 of the 20 urls (65%), while the remaining pages had been modified over the time course (mostly, for content updates).

For each url, we (automatically) retrieved the corresponding Web-page and formatted it as text by stripping off the HTML tags and other “boilerplate” (using Perl's HTML::TreeBuilder *as_text* function and simple regular expressions). Since Japanese pages can be in different character sets (e.g., shift-jis, euc-jp, iso-2022-jp, utf-8), our script extracts the character set in which a page is encoded from the HTML code, and converts from that character set into utf-8.

Since Japanese text does not use whitespace to separate word/characters, we used the ChaSen tool (Matsumoto et 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 (http://recode.progiciels-bpi.ca/) to convert back and forth between utf-8 and euc-jp.

According to the ChaSen tokenization, the 2004 corpus contains 3,539,109 tokens; the 2005 corpus contains 4,468,689 tokens. Thus, not only the repeated queries found different urls and more urls – they also found urls that contained more text. 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.

4. Corpus classification

For the qualitative evaluation of our automatically constructed corpora, the first author manually classified all 894 pages of the 2004 corpus and 300 randomly selected pages of the 2005 corpus in terms of topic domains and genre types.

4.1. Topic domain

For the classification o

f topics of the Web pages, we adopted the classification system proposed in Sharoff’s (to appear) study with minor modifications. We used the following ten categories:

***natscsi*:** agriculture, astronomy, meteorology, ...
***appsci*:** computing, engineering, transport, ...
***socsci*:** law, politics, sociology, language, education, religion...
***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 picked one in a coherent way. For example, we classified the Web pages 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.).

4.2. Genre type

Web pages are presented in various genre types, including the ones available in traditional corpora (e.g., news, diary...) and the ones newly emerging in internet use (e.g., blog). 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 first author first went through a good amount of the Web pages of the corpora to have a general idea of the distribution of genre types, and selected the following 24 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 site; 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 blog diaries 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 adapted 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 guide, guidelines, ...)
> ***info:*** pages that present information without a commercial/instructional purpose, (e.g., information on the time/place of a community event)
> ***teaching:*** materials for instruction, typically, language teaching (e.g., example sentences, language exercises, ...)
> ***news*****:** another genre type adapted from traditional genre types
> ***magazine:*** Web magazine
> ***report:*** report of academic research, report of events, ...
> ***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:*** list of links to Web pages with simple descriptions
> ***top:*** “top” pages that typically present the menu/structure of sites
> ***speech*****:** transcribed speeches

***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 *information* and *essay* into sub-categories depending on rhetorical types (i.e., argumentative, instructional...), being inspired in part by Santini (2005).

We originally used more than the 24 classes reported above, but for ease of post-classification analysis, we collapsed categories with less than 3 pages in either corpus into the *others* category.

5. Results

5.1. Domain

The downloaded Web pages were distributed across domains as shown in Table 1, where the number and percentage of documents and their average size in the number of tokens are summarized for each corpus:

	2004			***2005***		
	# of docs	*percentage*	*ave.size*	*# of docs*	*percentage*	*ave. size*
appsci	23	3%	2553	8	3%	1593
arts	37	4%	6820	13	4%	3295
business	216	24%	2516	51	17%	2170
error	47	5%	4522	18	6%	13396
leisure	188	21%	3753	65	22%	3653
life	283	32%	4579	111	37%	4564
natsci	15	2%	2843	5	2%	6078
socsci	86	10%	4268	29	10%	8323
total	894	100%	3888	300	100%	4744

Table 1: Distribution of domain types in the 2004 and 2005 corpora

Here we see that for both years *life, business, leisure* are the three major domain types among the Web pages of our corpora, although there is a difference in ranking: *life* > *business* > *leisure* in 2004; *life* > *leisure* > *business* in 2005, which suggests an increase in the proportion of “personal interest” pages.

Comparing our results with the ones of Sharoff’s (to appear) study (for corpora in English, Russian, German), we notice that the percentage of *socsci* is only about 10% in our corpora, while his corpora overall show higher percentages, ranging from 15% to 29% in the three languages (note that these are the sums of *politics* and *sosci,* which are collapsed in our classification system). Another difference concerns *life* and *leisure,* two domain types in which everyday life or personal interests are presented in Web pages. In our corpora, the sum of the two domain types is higher than 50% (53% in 2004, 59% in 2005). In Sharoff’s corpora, the value ranges from 25% (English) to 51% (Russian). We suspect that these differences between Sharoff’s corpora and our corpora are caused by 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 domains well-represented in these corpora that are also common on the

Web. It would be interesting to investigate the distribution of domain types in Web-based corpora for different languages that are constructed by using comparable seeds for all tested languages.

5.2. Genre types

The distribution of genre types is presented in the following table, which summarizes the number and the percentage of documents and their average size in the number of tokens for each genre type:

	2004			*2005*		
	# of docs	*percentage*	*ave.size*	*# of docs*	*percentage*	*ave. size*
BBS	54	6%	8243	10	3.3%	9329
QA	29	3.2%	2855	4	1.3%	2759
argessay	6	0.7%	3391	0	0.0%	0
blog	52	5.8%	4007	72	24.0%	4654
comments	10	1.1%	2040	10	3.3%	7936
commerinfo	159	17.8%	2399	40	13.3%	2305
diary	163	18.2%	5062	49	16.3%	5127
error	51	5.7%	4171	18	6.0%	13396
essay	66	7.4%	3414	15	5.0%	4082
info	80	8.9%	2700	22	7.3%	3607
instinfo	14	1.6%	4117	6	2.0%	2820
links	47	5.3%	1723	7	2.3%	2083
list	11	1.2%	2578	6	2.0%	550
magazine	11	1.2%	4779	0	0.0%	0
news	14	1.6%	3722	0	0.0%	0
novel	18	2.0%	10367	4	1.3%	3236
others	13	1.5%	6140	8	2.7%	7349
personal	16	1.8%	2138	3	1.0%	2077
questionnaire	21	2.3%	3878	5	1.7%	1393
report	31	3.5%	2587	11	3.7%	2429
review	5	0.6%	5733	0	0.0%	0
speech	4	0.4%	3561	4	1.3%	2671
teaching	7	0.8%	5854	3	1.0%	3741
top	12	1.3%	1706	3	1.0%	3659
total	894	100%	7776	300	100%	9488

Table 2: Distribution of genre types in the 2004 and 2005 corpora

The general pattern that we found here is that 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 39.2% in 2004 and 49.6% in 2005. The overall dominance of the personal genres indicates that the Web-based corpora are likely to include a good amount of spontaneous prose composed by non-professional writers. Since this type of prose is not available in traditional corpora, Web-based corpora can be a very precious linguistic resource. Interestingly, we notice a sharp increase in the overall proportion of these genres between 2004 and 2005, suggesting 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 17.8% and 13.3% of Web documents in the 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 57% and 62.9% in 2004 and 2005, respectively.

The ratio of *news* is surprisingly low (1.6% in 2004, 0% in 2005). This may be caused by our selection of seed terms, as was probably the case for the low percentage of *socsci* in the results of the domain evaluation presented in the previous section.

The genre types that do not include a good chunk of prose, such as *links* (links to other Web pages), *top (*top pages with a site menu) and *list* (lists of words or numbers), have a relatively low ratio (7.8% in 2004 and 5.3% in 2005 in total). This is, of course, a good thing.

The more common genres also show relatively stable average document size between the two corpora (e.g., *blog, commerinfo, diary*); while less represented genres show more fluctuation in size over the time course (e.g., *comments, instinfo, novel, list*). However, it is difficult to draw any sensible conclusion based on this observation, given that the standard deviations (not reported here) of all the genre types are very large.

In summary, the results of the genre classification show that a good majority of Web documents in our Japanese Web-based corpora are constituted by personal or commercial genres, which matches the results of the domain classification.

5.3. Typical lexical items in *blog* and *diary*

Our analysis showed that *blog* and *diary* are among the main genre types. How are these genre types characterized in terms of lexical items? Are there any differences or is this an arbitrary distinction? As an attempt to partly answer these questions, we conducted a qualitative analysis of typical lexical items of both genre types. This is just a small example case study illustrating a possible application of our corpora and their classification.

5.3.1. Data preparation

For both years, and for both the *blog* and *diary* genres, we compared the frequency of occurrence of each “word” (as tokenized by ChaSen) in the documents classified as belonging to the target genre with its frequency in all the other documents of the corpus by computing the log-likelihood ratio association measure (Dunning 1993). The first author then evaluated the lists of words ranked by log-likelihood ratio, focusing in particular on the top 300 items in each list.

5.3.2. Characteristic terms

Many terms in the top *blog* and *diary* lists cluster around specific topics/semantic fields, as shown in Table 3.

blog	**<*blog/internet*>** **トラックパック** *torakkupakku* “track pack,” **固定リンク** *koteerinku* “fixed link,” **ブログ** *burogu* “blog,” **ケータイ** *keetai* “cell phone,” **ソーシャルネットワーキング** *soosharunettowaakingu* “social networking,” **メールチェック** *meeruchekku* “mail check”
	<*loanwords*> **カフェコーナー** *kafekoonaa* “cafè corner,” **ショック** *shokku* “shock,” **ミゼラブル** mizeraburu “miserable,” **デジカメ** dezikame “digital camera,” **ハードディスク** haadodisuku “hard disk”
	<*trends/hot topics*> **サップ** *sappu* “Bob Sap (professional ring fighter),” **マツケン** *matstuken* “Matsudaira Ken (actor),” **個人情報流出** *kojinjyoohooryuushutsu* “personal information leak),” **ドンキ** *donki* “(abbreviation of) Don Quixote (discount chain store),” **ヨン様** *yonsama* “Mr. Yong (nick name of a very popular Korean actor),” **竹島** *takeshima* “Takeshima (island whose ownership is object of controversy between Korea and Japan)”
diary	**<*weather*>** **晴れ** *hare* “sunny weather,” **雨** *ame* “rain,” **曇り** *kumori* “cloudy weather”
	<*everyday life*> **会社** *kaisha* “company,” **帰宅** *kitaku* “coming back home,” **風呂** *furo* “bath,” **朝食** *chooshoku* “breakfast,” **夕食** *yuushoku* “dinner,” **お弁当** *obentoo* “box lunch”
	<*temporal expressions*> **明日** *ashita* “tomorrow,” **昨日** *kinoo* “yesterday,” **本日** *honjitsu* “today,” **午前中** *gozenchuu* “before noon,” **土曜日** *doyoobi* “Saturday”

Table 3: Characteristic terms in *blog* and *diary*

As exemplified in the table, *blog* is characterized by a strong presence of three kinds of terms: first, there are many blog-specific terms (e.g., **ブログ** *burogu* “blog,” **トラックパック** *torakkupakku* “track pack”) or new internet terms (e.g., **固定リンク** *koteerinku* “fixed link,” **ソーシャルネットワーキング** *soosharunettowaakingu* “social networking”); second, there is a high ratio of terms written in *katakana,* the writing system used especially to transcribe loanwords (e.g., **ショック** *shokku* “shock,” **ミゼラブル** mizeraburu “miserable”); third, there is a high ratio of words related to recently or currently hot topics in the Japanese society. Of the example *blogs* terms listed in Table 3, none is ranked higher than the 18,500th position in the *diary* list ranked by log-likelihood ratio.

In contrast, the characteristic items of *diary* pattern along the following three lexical types: first, weather terms (**晴れ** *hare* “sunny weather,” **雨** *ame* “rain,” **曇り** *kumori* “cloudy weather”); second, terms related to everyday life routines (e.g., **帰宅** *kitaku* “coming back home,” **風呂** *furo* “bath,” **夕食** *yuushoku* “dinner”); finally, temporal expressions (e.g., **明日** *ashita* “tomorrow,” **昨日** *kinoo* “yesterday,” **午前中** *gozenchuu* “before noon (A.M.),” **土曜日** *doyoobi* “Saturday”). None of the examples of *diary* in

Table 3 are ranked high in the log-likelihood ratio list of *blog* (none of them is within the top 18,000 *blog* terms). It is very interesting that these three lexical groups are not observed in *blog*, given that they seem fairly typical of diary prose, and a blog is arguably a form of diary. We see then that the difference between blogs and diaries emerge very clearly from the data. The *blog*, despite the similarities to diaries at least in terms of mode of production, is a novel genre that emerged recently for internet use, while *diary* is an genre directly adapted from the corresponding traditional paper genre (see Santini 2005).

5.3.3. Other differences

Other differences between the two genre types seem to pertain to register more than to the topical/lexical domain. First, the *blog* data, unlike the *diary* data, are characterized by the marked presence of onomatopoeia, which probably cues a more informal/oral-like register (see Shibatani 1990 for a description of onomatopoeia in Japanese).

A second difference is the use of more or less formal ways to refer to people. For example, *blogs* favor forms with the suffix *–chan*, used in less formal settings, whereas *diaries* are characterized by more forms with the more formal honorific marker *–san*. Similarly, *blogs* favor less formal, "younger generation" ways to refer to family members. Representative examples are shown in Table 4.

blog	***diary***
	連れ合いさん *tsureai-san* "company"
だんな／ダンナ *danna* "husband"	
嫁 *yome* "wife"	**妻** *tsuma* "wife" **カミさん** *kami-san* "wife" **奥さん** *oku-san* "wife"
お姉／お姉ちゃん *onee/onee-chan* "older sister"	
お兄 *onii* "older brother"	
お祖母ちゃん *obaa-chan* "granma"	**おじさん** *oji-san* "uncle"
知世ちゃん *tomiyo-chan* "Tomoyo (diminutive)""	**木村さん** *kimura-san* "Mr./Mrs. Kimura"

Table 4: Family member terms and person names in the top log-likelihood ratio lists

5.3.4. Summary

The analysis of the data ranked by log-likelihood ratio shows several differences in the distribution of frequent terms between *blog* and *diary*. First, *blog* is characterized by internet/blog terms, loanwords, and words related to trendy journalistic topics, while *diary* is characterized by words related to weather, daily routines and temporal references that are probably also typical of traditional paper diaries. Second, the data suggest that *blogs* are written in a less formal register than *diaries*.

What we reported here is true for both the 2004 and the 2005 corpus. However, there are also some intriguing differences emerging between the two years. For example, within the *blog* category, the prominence of internet/blog terms considerably decreased from 2004 to 2005, whereas that of trendy journalistic topics increased. We will not discuss this further here, but patterns such as the one we have just reported suggest that our method could be used to conduct short-term diachronic studies on the evolution of the language of the Web.

5. Conclusion

It should be clear that the experiments reported in this paper have a rather preliminary nature. Among the directions we intend to follow in further studies, we would like to test the effects of using seed terms of a different nature (e.g., extracted from newspapers), and to use the same type of seeds we employed here to collect an English corpus. Experiments with comparable seed sets in different languages should allow us to assess to what extent the textual types encountered on the Web for different languages (at least as seen through the filter of a search engine) are similar, and to what extent they differ. Also, by combining basic vocabulary and terms from other sources, we should probably be able to construct corpora that range over a very broad typology of genres and domains, something that should be a highly desirable property for reference corpora.

In the meantime, our current results indicate that the automated search engine query method we illustrated can be used to construct corpora that are reasonably clean and varied. The presence of many documents produced by non-professional writers, characterized by everyday topics and by an often informal, spontaneous, interactive style should make corpora constructed in this way a very precious resource for linguistic research, and suggests that Web data are at least in part complementary to those encountered in traditional corpora

Our comparison of data collected in the same way in 2004 and 2005, on one hand, shows that the methodology could be used to detect short-term diachronic trends (which, of course, will pertain more on what is popular on the Web, than on the deep characteristics of a language). On the other hand, the major changes in retrieved pages shows very clearly that queries made at different times to the same engine cannot be treated as queries to the same corpus. Studies based on Web data will be truly comparable only if they use the same set of retrieved pages.

References

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

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

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, W. (2004) Making the Web more useful as a source for linguistic corpora, in Cornnor, U. and Upton, T. (eds.) *Corpus Linguistics in Nort 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/

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

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

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

Meyer zu Eissen, S. and Stein, B. (2004) Genre classification of Web pages, *KI-2004*, 256-269.

Santini, M. (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/

Shibatani, M. (1990) *The Languages of Japan.* (Cambridge: Cambridge University Press).

Sharoff, S. (to appear). Open-source corpora: using the net to fish for linguistic data. To appear in the *International Journal of Corpus Linguistics*.