A Figure of Merit for the Evaluation of Web-Corpus Randomness

Massimiliano Ciaramita
Institute of Cognitive Science and Technology
National Research Council
Roma, Italy
m.ciaramita@istc.cnr.it

Marco Baroni
SSLMIT
Università di Bologna
Forlì, Italy
baroni@sslmit.unibo.it

Abstract

In this paper, we present an automated, quantitative, knowledge-poor method to evaluate the randomness of a collection of documents (corpus), with respect to a number of biased partitions. The method is based on the comparison of the word frequency distribution of the target corpus to word frequency distributions from corpora built in deliberately biased ways. We apply the method to the task of building a corpus via queries to Google. Our results indicate that this approach can be used, reliably, to discriminate biased and unbiased document collections and to choose the most appropriate query terms.

1 Introduction

The Web is a very rich source of linguistic data, and in the last few years it has been used intensively by linguists and language technologists for many tasks (Kilgarriff and Grefenstette, 2003). Among other uses, the Web allows fast and inexpensive construction of "general purpose" corpora, i.e., corpora that are not meant to represent a specific sub-language, but a language as a whole. There are several recent studies on the extent to which Web-derived corpora are comparable, in terms of variety of topics and styles, to traditional "balanced" corpora (Fletcher, 2004; Sharoff, 2006). Our contribution, in this paper, is to present an automated, quantitative method to evaluate the "variety" or "randomness" (with respect to a number of non-random partitions) of a Web corpus. The more random/less-biased towards specific partitions a corpus is, the more it should be suitable as a general purpose corpus.

We are not proposing a method to evaluate whether a sample of Web pages is a random sample of the Web, although this is a related issue (Bharat and Broder, 1998; Henzinger et al., 2000). Instead, we propose a method, based on simple distributional properties, to evaluate if a sample of Web pages in a certain language is reasonably varied in terms of the topics (and, perhaps, textual types) it contains. This is independent from whether they are actually proportionally representing what is out there on the Web or not. For example, although computer-related technical language is probably much more common on the Web than, say, the language of literary criticism, one might prefer a biased retrieval method that fetches documents representing these and other sub-languages in comparable amounts, to an unbiased method that leads to a corpus composed mostly of computer jargon. This is a new area of investigation – with traditional corpora, one knows *a priori* their composition. As the Web plays an increasingly central role as data source in NLP, we believe that methods to efficiently characterize the nature of automatically retrieved data are becoming of central importance to the discipline.

In the empirical evaluation of the method, we focus on general purpose corpora built issuing automated queries to a search engine and retrieving the corresponding pages, which has been shown to be an easy and effective way to build Web-based corpora (Ghani et al., 2001; Ueyama and Baroni, 2005; Sharoff, 2006). It is natural to ask which kinds of query terms, henceforth *seeds*, are more appropriate to build a corpus comparable, in terms of variety, to traditional balanced corpora such as the British National Corpus, henceforth BNC (Aston and Burnard, 1998). We test our procedure to assess Web-corpus randomness on corpora built

using seeds chosen following different strategies. However, the method *per se* can also be used to assess the randomness of corpora built in other ways; e.g., by crawling the Web.

Our method is based on the comparison of the word frequency distribution of the target corpus to word frequency distributions constructed using queries to a search engine for deliberately biased seeds. As such, it is nearly resource-free, as it only requires lists of words belonging to specific domains that can be used as biased seeds. In our experiments we used Google as the search engine of choice, but different search engines could be used as well, or other ways to obtain collections of biased documents, e.g., via a directory of pre-categorized Web-pages.

2 Relevant work

Our work is related to the recent literature on building linguistic corpora from the Web using automated queries to search engines (Ghani et al., 2001; Fletcher, 2004; Ueyama and Baroni, 2005; Sharoff, 2006). Different criteria are used to select the seeds. Ghani and colleagues iteratively bootstrapped queries to AltaVista from retrieved documents in the target language and in other languages. They seeded the bootstrap procedure with manually selected documents, or with small sets of words provided by native speakers of the language. They showed that the procedure produces a corpus that contains, mostly, pages in the relevant language, but they did not evaluate the results in terms of quality or variety. Fletcher (2004) constructed a corpus of English by querying AltaVista for the 10 top frequency words from the BNC. He then conducted a qualitative analysis of frequent n-grams in the Web corpus and in the BNC, highlighting the differences between the two corpora. Sharoff (2006) built corpora of English, Russian and German via queries to Google seeded with manually cleaned lists of words that are frequent in a reference corpus in the relevant language, excluding function words, while Ueyama and Baroni (2005) built corpora of Japanese using seed words from a basic Japanese vocabulary list. Both Sharoff and Ueyama and Baroni evaluated the results through a manual classification of the retrieved pages and by qualitative analysis of the words that are most typical of the Web corpora.

We are also interested in evaluating the effect that different seed selection (or, more in general, corpus building) strategies have on the nature of the resulting Web corpus. However, rather than performing a qualitative investigation, we develop a quantitative measure that could be used to evaluate and compare a large number of different corpus building methods, as it does not require manual intervention. Moreover, our emphasis is not on the corpus building methodology, nor on classifying the retrieved pages, but on assessing whether they appear to be reasonably unbiased with respect to a range of topics or other criteria.

3 Measuring distributional properties of biased and unbiased collections

Our goal is to create a "balanced" corpus of Web pages in a given language; e.g., the portion composed of all Spanish Web pages. As we observed in the introduction, obtaining a sample of unbiased documents is not the same as obtaining an unbiased sample of documents. Thus, we will not motivate our method in terms of whether it favors unbiased samples from the Web, but in terms of whether the documents that are sampled appear to be balanced with respect to a set of deliberately biased samples. We leave it to further research to investigate how the choice of the biased sampling method affects the performance of our procedure and its relations to uniform sampling.

3.1 Corpora as unigram distributions

A compact way of representing a collection of documents is by means of frequency lists, where each word is associated with the number of times it occurs in the collection. This representation defines a simple "language model", a stochastic approximation to the language of the collection; i.e., a "0th order" word model or a "unigram" model. Language models of varying complexity can be defined. As the model's complexity increases, its approximation to the target language improves – cf. the classic example of Shannon (1948) on the entropy of English. In this paper we focus on unigram models, as a natural starting point, however the approach extends naturally to more complex language models.

3.2 Corpus similarity measure

We start by making the assumption that similar collections will determine similar language models, hence that the similarity of collections of documents is closely related to the similarity of the

derived unigram distributions. The similarity of two unigram distributions P and Q is estimated as the *relative entropy*, or *Kullback Leibler distance*, or KL (Cover and Thomas, 1991) $D(P||Q)$:

$$D(P||Q) = \sum_{x \in W} P(x) \log \frac{P(x)}{Q(x)} \quad (1)$$

KL is a measure of the cost, in terms of average number of additional bits needed to describe the random variable, of assuming that the distribution is Q when instead the true distribution is P. Since $D(P||Q) \geq 0$, with equality only if $P = Q$, unigram distributions generated by similar collections should have low relative entropy. To guarantee that KL is always finite we make the assumption that the random variables are defined over the same finite alphabet W, the set of all word types occurring in the observed data. To avoid further infinite cases a smoothing value α is added when estimating probabilities; i.e.,

$$P(x) = \frac{c_P(x) + \alpha}{|W|\alpha + \sum_{x \in W} c_P(x)} \quad (2)$$

where $c_P(x)$ is the frequency of x in distribution P, and $|W|$ is the number of word types in W.

3.3 A scoring function for sampled unigram distributions

What properties distinguish unigram distributions drawn from the whole of a document collection such as the BNC or the Web (or, rather, from the space of the Web we are interested in sampling from) from distributions drawn from biased subsets of it? This is an important question because, if identified, such properties might help discriminating between sampling methods which produce more random collections of documents from more biased ones. We suggest the following hypothesis. Unigrams sampled from the full set of documents have distances from biased samples which tend to be lower than the distances of biased samples to other samples based on different biases. Samples from the whole corpus, or Web, should produce lower KL distances because they draw words across the whole vocabulary, while biased samples have mostly access to a single specialized vocabulary. If this hypothesis is true then, on average, the distance between the unbiased sample and all other samples should be lower than the distance between a biased sample and all other samples.

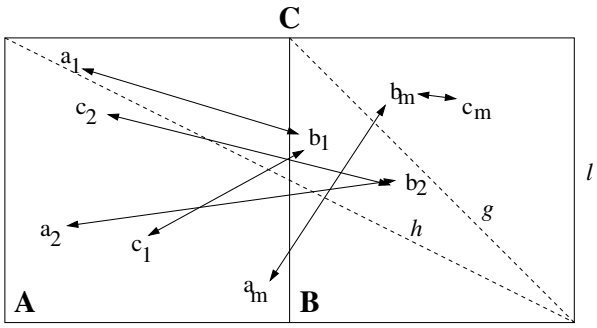

Figure 1. Distances (continuous lines with arrows) between points representing unigram distributions, sampled from biased partitions A and B and from the full collection of documents $C = A \cup B$.

Figure 1 depicts a geometric interpretation of the intuition behind this hypothesis. Suppose that the two squares A and B represent two partitions of the space of documents C. Additionally, m pairs of unigram distributions, represented as points, are produced by sampling documents uniformly at random from these partitions; e.g. a_1 and b_1. The mean Euclidean distance between (a_i, b_i) pairs is a value between 0 and h, the length of the diagonal of the rectangle which is the union of A and B. Instead of drawing pairs we can draw triples of points, one point from A, one from B, and another point from $C = A \cup B$. Approximately half of the points drawn from C will lie in the A square, while the other half will lie in the B square. The distance of the points drawn from C from the points drawn from B will be between 0 and g, for approximately half of the points (those laying in the B region), while the distance is between 0 and h for the other half of the points (those in A). Therefore, if m is large enough, the average distance between C and B (or A) must be smaller than the average distance between A and B, because $h > g$.

To summarize, then, we suggest the hypothesis that samples from the full distribution have a smaller mean distance than all other samples. More precisely, let $U_{i,k}$ be the kth of N unigram distributions sampled with method y_i, $y_i \in Y$, where Y is the set of sampling categories. Additionally, for clarity, we will always denote with y_1 the predicted unbiased sample, while y_j, $j = 2..|Y|$, denote the biased samples. Let $\mathbf{M}$ be a matrix of measurements, $\mathbf{M} \in \mathbb{R}^{|Y| \times |Y|}$, such that $M_{i.j} = \frac{\sum_{k=1}^{N} D(U_{i,k}, U_{j,k})}{N}$, where $D(.,.)$ is the relative entropy. In other words, the matrix contains the average distances between pairs of sam-

	Mode	Domain	Genre
1	BNC	BNC	BNC
2	W	S_education	W_miscellaneous
3	S	W_leisure	W_pop_lore
4		W_arts	W_nonacad_soc_sci
5		W_belief_thought	W_nonacad_hum_art
..		..	..
C-4		S_spont_conv_C1	S_sportslive
C-3		S_spont_conv_C2	S_consultation
C-2		S_spont_conv_DE	W_fict_drama
C-1		S_spont_conv_UN	S_lect_commerce
C		no_cat	no_cat

Table 1. Rankings based on δ, as the mean distance between samples from the BNC partitions plus samples from the whole corpus (BNC). C is the total number of categories. W stands for Written, S for Spoken. C1, C2, DE, UN are demographic classes for the spontaneous conversations, no_cat is the BNC undefined category.

ples (biased or unbiased). Each row $M_i \in \mathbb{R}^{|Y|}$ contains the average distances between y_i and all other ys, including y_i. A score δ_i is assigned to each y_i which is equal to the mean of the vector M_i (excluding $M_{i,j}$, $j = i$, which is always equal to 0):

$$\delta_i = \frac{1}{|Y| - 1} \sum_{j=1, j \neq i}^{|Y|} M_{i,j} \quad (3)$$

We propose this function as a *figure of merit*[1] for assigning a score to sampling methods. The smaller the δ value the closer the sampling method is to a uniform sampling method, with respect to the pre-defined set of biased sampling categories.

3.4 Randomness of BNC samples

Later we will show how this hypothesis is consistent with empirical evidence gathered from Web data. Here we illustrate a proof-of-concept experiment conducted on the BNC. In the BNC documents come classified along different dimensions thus providing a controlled environment to test our hypothesis. We adopt here David Lee's revised classification (Lee, 2001) and we partition the documents in terms of "mode" (spoken/written), "domain" (19 labels; e.g., imaginative, leisure, etc.) and "genre" (71 labels; e.g., interview, advertisement, email, etc.). For each of the three main partitions we sampled with replacement (from a distribution determined by relative frequency in the relevant set) 1,000 words from the BNC and from each of the labels belonging to the specific partitions.[2] Then we measured the distance between each label in a partition, plus the sample from the whole BNC. We repeated this experiment 100 times, built a matrix of average distances, and ranked each label y_i, within each partition type, using δ_i. Table 1 summarizes the results (only partial results are shown for domain and genre). In all three experiments the unbiased sample "BNC" is ranked higher than all other categories. At the top of the rankings we also find other less narrowly topic/genre-dependent categories such as "W" for mode, or "W_miscellaneous" and "W_pop_lore" for genre. Thus the hypothesis seems supported by these experiments. Unbiased sampled unigrams tend to be closer, on average, to biased samples.

4 Evaluating the randomness of Google-derived corpora

When downloading documents from the Web via a search engine (or sample them in other ways), one cannot choose to sample randomly, nor select documents belonging to a certain category. One can try to control the typology of documents returned by using specific query terms. At this point a measure such as the one we proposed can be used to choose the least biased retrieved collection among a set of retrieved collections.

4.1 Biased and unbiased query categories

To construct a "balanced" corpus via a search engine one reasonable strategy is to use appropriately balanced query terms, e.g., using random terms extracted from an available balanced corpus (Sharoff, 2006). We will evaluate several such strategies by comparing the derived collections with those obtained with openly biased/specialized Web corpora. In order to build specialized domain corpora, we use biased query terms from the appropriate domain following the approach of Baroni and Bernardini (2004). We compiled several lists of words that define likely biased and unbiased categories. We extracted the less biased terms from the balanced 1M-words Brown corpus of American English (Kučera and Francis, 1967), from the 100M-words BNC, and from a list of English "basic" terms. From these resources we defined the following categories of query terms:

[1] A function which measures the quality of the sampling method with the convention that smaller values are better as with merit functions in statistics.

[2] We filtered out words in a stop list containing 1,430 types, which were either labeled with one of the BNC function word tags (such as "article" or "coordinating conjunction"), or occurred more than 50,000 times.

1. **Brown.hf**: the top 200 most frequent words from the Brown corpus;

2. **Brown.mf**: 200 random terms with frequency between 100 and 50 inclusive from Brown;

3. **Brown.af**: 200 random terms with minimum frequency 10 from Brown;

4. **BNC.mf**: 200 random terms with frequency between 506 and 104 inclusive from BNC;

5. **BNC.af**: 200 random terms from BNC;

6. **BNC.demog**: 200 random terms with frequency between 1000 and 50 inclusive from the BNC spontaneous conversation sections;

7. **3esl**: 200 random terms from an ESL "core vocabulary" list.[3]

Some of these lists implement plausible strategies to get an unbiased sample from the search engine: high frequency words and basic vocabulary words should not be linked to any specific domain; while medium frequency words, such as the words in the Brown.mf/af and BNC.mf lists, should be spread across a variety of domains and styles. The BNC.af list is sampled randomly from the whole BNC and, because of the Zipfian properties of word types, coupled with the large size of the BNC, it is mostly characterized by very low frequency words. In this case, we might expect data sparseness problems. Finally, we expect the spoken demographic sample to be a "mildly biased" set, as it samples only words used in spoken conversational English.

In order to build biased queries, hopefully leading to the retrieval of topically related documents, we defined a set of specialized categories using the WordNet (Fellbaum, 1998) "domain" lists (Magnini and Cavaglia, 2000). We selected 200 words at random from each of the following domains: *administration, commerce, computer science, fashion, gastronomy, geography, law, military, music, sociology*. These domains were chosen since they look "general" enough that they should be very well-represented on the Web, but not so general as to be virtually unbiased (cf. the WordNet domain *person*). We selected words only among those that did not belong to more than one WordNet domain, and we avoided multi-word terms.

It is important to realize that a balanced corpus is not necessary to produce unbiased seeds, nor a topic-annotated lexical resource for biased seeds. Here we focus on these sources to test plausible candidate seeds. However, biased seeds can be obtained following the method of Baroni and Bernardini (2004) for building specialized corpora, while unbiased seeds could be selected, for example, from word lists extracted from all corpora obtained using the biased seeds.

4.2 Experimental setting

From each source list we randomly select 20 pairs of words without replacement. Each pair is used as a query to Google, asking for pages in English only. Pairs are used instead of single words to maximize our chances to find documents that contain running text (Sharoff, 2006). For each query, we retrieve a maximum of 20 documents. The whole procedure is repeated 20 times with all lists, so that we can compute the mean distances to fill the distance matrices. Our unit of analysis is the corpus of all the non-duplicated documents retrieved with a set of 20 paired word queries. The documents retrieved from the Web undergo post-processing, including filtering by minimum and maximum size, removal of HTML code and "boilerplate" (navigational information and similar) and heuristic filtering of documents that do not contain connected text. A corpus can contain maximally 400 documents (20 queries times 20 documents retrieved per query), although typically the documents retrieved are less, because of duplicates, or because some query pairs are found in less than 20 documents. Table 2 summarizes the average size in terms of word types, tokens and number of documents of the resulting corpora. Queries for the unbiased seeds tend to retrieve more documents except for the BNC.af set, which, as expected, found considerably less data than the other unbiased sets. Most of the differences are not statistically significant and, as the table shows, the difference in number of documents is often counterbalanced by the fact that specialized queries tend to retrieve longer documents.

4.3 Distance matrices and bootstrap error estimation

After collecting the data each sample was represented as a frequency list as we did before with

[3] http://wordlist.sourceforge.net/12dicts-readme.html

Search category	Types	Tokens	Docs
Brown.hf	39.3	477.2	277.2
Brown.mf	32.8	385.3	261.1
Brown.af	35.9	441.5	262.5
BNC.mf	45.6	614.7	253.6
BNC.af	23.0	241.7	59.7
BNC.demog	32.6	367.1	232.2
3esl	47.1	653.2	261.9
Admin	39.8	545.1	220.5
Commerce	38.9	464.5	184.7
Comp_sci	25.8	311.5	185.3
Fashion	44.5	533.7	166.2
Gastronomy	36.5	421.7	159.0
Geography	42.7	498.0	167;6
Law	49.2	745.4	211.4
Military	47.1	667.8	223.0
Music	45.5	558.7	201.3
Sociology	56.0	959.5	258.8

Table 2. Average number of types, tokens and documents of corpora constructed with Google queries (type and token sizes in thousands).

the BNC partitions (cf. section 3.4). Unigram distributions resulting from different search strategies were compared by building a matrix of mean distances between pairs of unigram distributions. Rows and columns of the matrices are indexed by the query category, the first category corresponds to one unbiased query, while the remaining indexes correspond to the biased query categories; i.e., $M \in \mathbb{R}^{11\times 11}$, $M_{i,j} = \frac{\sum_{k=1}^{20} D(U_{i,k},U_{j,k})}{20}$, where $U_{s,k}$ is the kth unigram distribution produced with query category y_s.

These Web-corpora can be seen as a dataset $\mathcal{D}$ of $n = 20$ data-points each consisting of a series of unigram word distributions, one for each search category. If all n data-points are used once to build the distance matrix we obtain one such matrix for each unbiased category and rank each search strategy y_i using δ_i, as before (cf. section 3.3). Instead of using all n data-points once, we create B "bootstrap" datasets (Duda et al., 2001) by randomly selecting n data-points from $\mathcal{D}$ with replacement (we used a value of B=10). The B bootstrap datasets are treated as independent sets and used to produce B individual matrices M_b from which we compute the score $\delta_{i,b}$, i.e., the mean distance of a category y_i with respect to all other query categories in that specific bootstrap dataset. The bootstrap estimate of δ_i, called $\hat{\delta}_i$ is the mean of the B estimates on the individual datasets:

$$\hat{\delta}_i = \frac{1}{B}\sum_{b=1}^{B}\delta_{i,b} \tag{4}$$

Bootstrap estimation can be used to compute the standard error of δ_i:

$$\sigma_{boot}[\delta_i] = \sqrt{\frac{1}{B}\sum_{b=1}^{B}[\hat{\delta}_i - \delta_{i,b}]^2} \tag{5}$$

Instead of building one matrix of average distances over N trials, we could build N matrices and compute the variance from there rather than with bootstrap methods. However this second methodology produces noisier results. The reason for this is that our hypothesis rests on the assumption that the estimated average distance is reliable. Otherwise, the distance of two arbitrary biased distributions can very well be smaller than the distance of one unbiased and a biased one, producing noisier measurements.

As we did before for the BNC data, we smoothed the word counts by adding a count of 1 to all words in the overall dictionary. This dictionary is approximated with the set of all words occurring in the unigrams involved in a given experiment, overall on average approximately 1.8 million types (notice that numbers and other special tokens are boosting up this total). Words with an overall frequency greater than 50,000 are treated as stop words and excluded from consideration (188 types).

5 Results

Table 3 summarizes the results of the experiments with Google. Each column represents one experiment involving a specific – supposedly – unbiased category. The category with the best (lowest) δ score is highlighted in bold. The unbiased sample is always ranked higher than all biased samples. The results show that the best results are achieved with Brown corpus seeds. The bootstrapped error estimate shows that the unbiased Brown samples are significantly more random than the biased samples and, orthogonally, of the BNC and 3esl samples. In particular medium frequency terms seem to produce the best results, although the difference among the three Brown categories are not significant. Thus, while more testing is needed, our data provide some support for the choice of medium frequency words as best seeds.

Terms extracted from the BNC are less effective than terms from the Brown corpus. One possible explanation is that the Web is likely to contain much larger portions of American than British English, and thus the BNC queries are overall

	δ scores with bootstrap error estimates						
Category	Brown.mf	Brown.af	Brown.hf	BNC.mf	BNC.demog	BNC.all	3esl
Unbiased	**.1248/.0015**	**.1307/.0019**	**.1314/.0010**	**.1569/.0025**	**.1616/.0026**	**.1635/.0026**	**.1668/.0030**
Commerce	.1500/.0074	.1500/.0074	.1500/.0073	.1708/.0088	.1756/.0090	.1771/.0091	.1829/.0093
Geography	.1702/.0084	.1702/.0084	.1707/.0083	.1925/.0089	.1977/.0091	.1994/.0092	.2059/.0094
Fashion	.1732/.0060	.1732/.0060	.1733/.0059	.1949/.0069	.2002/.0070	.2019/.0071	.2087/.0073
Admin	.1738/.0034	.1738/.0034	.1738/.0033	.2023/.0037	.2079/.0038	.2096/.0038	.2163/.0039
Comp_sci	.1749/.0037	.1749/.0037	.1746/.0038	.1858/.0041	.1912/.0042	.1929/.0042	.1995/.0043
Military	.1899/.0070	.1899/.0070	.1901/.0067	.2233/.0079	.2291/.0081	.2311/.0082	.2384/.0084
Music	.1959/.0067	.1959/.0067	.1962/.0067	.2196/.0077	.2255/.0078	.2274/.0079	.2347/.0081
Gastronomy	.1973/.0122	.1973/.0122	.1981/.0120	.2116/.0133	.2116/.0133	.2193/.0138	.2266/.0142
Law	.1997/.0060	.1997/.0060	.1990/.0061	.2373/.0067	.2435/.0068	.2193/.0138	.2533/.0070
Sociology	.2393/.0063	.2393/.0063	.2389/.0062	.2885/.0069	.2956/.0070	.2980/.0071	.3071/.0073

Table 3. Mean scores based on δ with bootstrap standard error (B=10). In bold the lowest (best) score in each column, always the unbiased category.

more biased than the Brown queries. Alternatively, this might be due to the smaller, more controlled nature of the Brown corpus, where even medium- and low-frequency words tend to be relatively common terms. The internal ranking of the BNC categories, although not statistically significant, seems also to suggest that medium frequency words (BNC.mf) are better than low frequency words. In this case, the all/low frequency set (BNC.af) tends to contain very infrequent words; thus, the poor performance is likely due to data sparseness issues, as also indicated by the relatively smaller quantity of data retrieved (Table 2 above). We take the comparatively lower rank of BNC.demog to constitute further support for the validity of our method, given that the corresponding set, being entirely composed of words from spoken English, should be more biased than other unbiased sets. This latter finding is particularly encouraging because the way in which this set is biased, i.e., in terms of mode of communication, is completely different from the topic-based bias of the WordNet sets. Finally, the queries extracted from the 3esl set are the most biased. This unexpected result might relate to the fact that, on a quick inspection, many words in this set, far from being what we would intuitively consider "core" vocabulary, are rather cultivated, often technical terms (*aesthetics*, *octopi*, *misjudgment*, *hydroplane*), and thus they might show a register-based bias that we do not find in lists extracted from balanced corpora. We randomly selected 100 documents from the corpora constructed with the "best" unbiased set (Brown.mf) and 100 documents from this set, and we classified them in terms of genre, topic and other categories (in random order, so that the source of the rated documents was not known). This preliminary analysis did not highlight dramatic differences between the two corpora, except for the fact that 6 over 100 documents in the 3esl subcorpus pertained to the rather narrow domain of aviation and space travel, while no comparably narrow topic had such a large share of the distribution in the Brown.mf sub-corpus. More research is needed into the qualitative differences that correlate with our figure of merit. Finally, although different query sets retrieve different amounts of documents, and lead to the construction of corpora of different lengths, there is no sign that these differences are affecting our figure of merit in a systematic way; e.g., some of the larger collections, in terms of number of documents and token size, are both at the top (most unbiased samples) and at the bottom of the ranks (law, sociology).

On Web data we observed the same effect we saw with the BNC data, where we could directly sample from the whole collection and from its biased partitions. This provides support for the hypothesis that our measure can be used to evaluate how unbiased a corpus is, and that issuing unbiased/biased queries to a search engine is a viable, nearly knowledge-free way to create unbiased corpora, and biased corpora to compare them against.

6 Conclusion

As research based on the Web as corpus becomes more prominent within computational and corpus-based linguistics, many fundamental issues have to be tackled in a systematic way. Among these, the problem of assessing the quality and nature of automatically created corpora, where we do not know *a priori* the composition of the corpus. In this paper, we considered an approach to automated corpus construction, via search engine queries for combinations of a set of seed words.

We proposed an automated, quantitative, nearly knowledge-free way to evaluate how biased a corpus constructed in this way is. Our method is based on the idea that the more a collection is unbiased the closer its distribution of words will be, on average, to reference distributions derived from biased partitions (we showed that this is indeed the case using a fully available balanced collection; i.e., the BNC), and on the idea that biased collections of Web documents can be created by issuing biased queries to a search engine. The results of our experiments with Google support our hypothesis, and suggest that seeds to build unbiased corpora should be selected among mid-frequency words rather than high or low frequency words.

We realize that our study opens many questions. The most crucial issue is probably what it means for a corpus to be unbiased. As we already stressed, we do not necessarily want our corpus to be an unbiased sample of what is out there on the Net – we want it to be composed of content-rich pages, and reasonably balanced in terms of topics and genres, despite the fact that the Web itself is unlikely to be "balanced". For our purposes, we implicitly define balance in terms of the set of biased corpora that we compare the target corpus against. Assuming that our measure is appropriate, what it tells us is that a certain corpus is more/less biased than another corpus with respect to the biased corpora they are compared against. It remains to be seen how well the results generalize across different typologies of biased corpora.

The method is not limited to the evaluation of corpora built via search engine queries; e.g., it would be interesting to compare the latter to corpora built by Web crawling. The method could be also applied to the analysis of corpora in general (Web-derived or not), both for the purpose of evaluating biased-ness, and as a general purpose corpus comparison technique (Kilgarriff, 2001).

Acknowledgments

We would like to thank Ioannis Kontoyiannis, Adam Kilgarriff and Silvia Bernardini for useful comments on this work.

References

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

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC 2004*, pages 1313–1316.

K. Bharat and A. Broder. 1998. A Technique for Measuring the Relative Size and Overlap of the Public Web Search Engines. In *Proceedings of WWW7*, pages 379–388.

T.M. Cover and J.A. Thomas. 1991. *Elements of Information Theory*. Wiley, New York.

R.O. Duda, P.E. Hart, and D.G. Stork. 2001. *Pattern Classification 2nd ed.* Wiley Interscience, Wiley Interscience.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

B. Fletcher. 2004. Making the Web more Useful as a Source for Linguistic Corpora. In U. Conor and T. Upton, editors, *Corpus Linguistics in North America 2002*. Rodopi, Amsterdam.

R. Ghani, R. Jones, and D. Mladenic. 2001. Using the Web to Create Minority Language Corpora. In *Proceedings of the 10th International Conference on Information and Knowledge Management*.

M. Henzinger, A. Heydon, and M. Najork. 2000. On Near-Uniform URL Sampling. In *Proceedings of WWW9*.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29:333–347.

A. Kilgarriff. 2001. Comparing Corpora. *International Journal of Corpus Linguistics*, 6:1–37.

H. Kučera and W. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

D. Lee. 2001. Genres, Registers, Text, Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology*, 5(3):37–72.

B. Magnini and G. Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC 2000, Athens*, pages 1413–1418.

C.E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

S. Sharoff. 2006. Creating General-Purpose Corpora Using Automated Search Engine Queries. In M. Baroni and S. Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

M. Ueyama and M. Baroni. 2005. Automated Construction and Evaluation of a Japanese Web-Based Reference Corpus. In *Proceedings of Corpus Linguistics 2005*.