Measuring Web Corpus Randomness: A Progress Report

Massimiliano Ciaramita and Marco Baroni

1 Introduction

The Web is a very rich source of linguistic data, and in the last few years it has been used very intensively by linguists and language technologists for many tasks (see Kilgarriff and Grefenstette 2003 for a review of some of the relevant work). Among other uses, the Web allows fast and inexpensive construction of "reference"/"general-purpose" corpora, i.e., corpora that are not meant to represent a specific sub-language, but a language as a whole. There is a large literature on the issue of representativeness of corpora (see, e.g., Biber 1993), and 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 (e.g., Fletcher 2004, Sharoff this volume). 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 a specific partition a corpus is, the more it should be suitable as a general-purpose corpus. It is important to realize that we are not proposing a method to evaluate whether a sample of webpages is a random sample of the Web. Instead, we are proposing a method to evaluate if a sample of webpages in a certain language is reasonably varied in terms of the topics (and, perhaps, textual types) it represents.

In our 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 (see section 2 below). With respect to this approach, it is natural to ask which kinds of query terms (henceforth seeds) are more appropriate to build a corpus that is comparable, in terms of variety and representativeness, to a traditional balanced corpus such as the British National Corpus (BNC). We will test our method for assessing Web corpus randomness on corpora built with low, medium and high frequency seeds. 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 starting from a few selected URLs).

Our method is based on the comparison of the word frequency distributions of the target corpus to word frequency distributions constructed using queries to a search engine for deliberately biased seeds (i.e., instead of trying to compare the corpus to a supposedly unbiased corpus, we look at how it compares to corpora that we are almost certain are highly biased). 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. While in our experiments we used Google as the search engine of choice, and in what follows we often use "Google" and "search engine" interchangeably, our procedure could also be carried out using a different search engine (or other ways to obtain collections of biased documents, e.g., via a directory of pre-categorized webpages).

After reviewing some of the relevant literature in section 2, in section 3 we introduce and justify our methodology. We show how, when we can sample randomly from the whole BNC and from its domain and genre partitions, our method to measure distance between sets of documents produces intuitively plausible results (similar partitions are nearer each other), and that the most varied, least biased distribution (the one from the whole BNC) is the one that has the least average distance from all the other (biased) distributions (we provide a geometric explanation of why this is the case). Hence, we propose average distance from a set of biased distributions as a way to measure corpus randomness: the lower the average distance, the more random/unbiased the corpus is. In section 4, we apply our technique to unbiased and biased corpora constructed via Google queries. The results of the Google experiments are very encouraging, in that the corpora built with various unbiased seed sets show, systematically, significantly shorter average distance to the biased corpora than any corpus built with biased seeds. Among unbiased seed sets chosen from high and medium frequency words, and from the whole frequency range, medium frequency words appear to be the best (in the sense that they lead to the least biased corpus, according to our method). In section 5, we conclude by summarizing our main results, considering some open questions and sketching directions for further work.

2 Relevant work

Our work is obviously related to the recent literature on building linguistic corpora from the Web using automated queries to search engines (see, e.g., Ghani et al. 2001, Fletcher 2004, Baroni and Bernardini 2004, Sharoff this volume, Ueyama this volume). With the exception of Baroni and Bernardini, who are interested in the construction of specialized language corpora, these researchers use the technique to build corpora that are meant to function as general-purpose "reference" corpora for the relevant language.

Different criteria are used to select seed words. Ghani and colleagues iteratively bootstrap queries to AltaVista from retrieved documents in the target language and in other languages. They seed the bootstrap procedure with manually selected documents, or with small sets of words provided by native speakers of the target language. They evaluate performance in terms of how many of the retrieved pages are in the relevant language, but do not assess their quality or variety. Fletcher constructs a corpus of English by querying AltaVista for the 10 top frequency words from the BNC. He then conducts a qualitative analysis of frequent n-grams in the Web corpus and in the BNC, highlighting the differences between the two corpora. Sharoff (this volume; see also Sharoff submitted) builds corpora of English, Russian and German using queries to the Google search engine, seeded with manually cleaned lists of words that are frequent in a reference corpus in the relevant language, excluding function words. Sharoff evaluates the results both in terms of manual classification of the retrieved pages and by means of a qualitative analysis of the words that are most typical of Web corpora vs. other corpora. For English, he also provides a comparison of corpora retrieved using non-overlapping but similarly selected seed sets, concluding that the difference in seeds is not having a strong effect on the nature of the pages retrieved. Ueyama (this volume; see also Ueyama and Baroni 2005) builds corpora of Japanese using both words from a basic Japanese vocabulary list, and translations from one of Sharoff's English lists (based on the BNC) as seeds. Through qualitative methods similar to those of Sharoff, she shows how the corpus built using basic vocabulary seeds is characterized by more "personal" genres than the one constructed from BNC-style seeds.

Like Sharoff and Ueyama, we are 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.

A different line of research somewhat related to ours pertains to the development of methods to perform quasi-random sampling of documents from the Web. The emphasis is not on corpus building, but on estimating statistics such as the percentage of pages in a certain domain, or the size and overlap of pages indexed by different search engines. For example, both Henzinger et al. (2000) and Bar-Yossef et al. (2000) use random walks through the Web, represented as a graph, to answer questions of this kind. Bharat and Broder (1998) issue random queries (based on words extracted from documents in the Yahoo! hierarchy) to various search engines in order to estimate their relative size and overlap. There are two important differences between work in this tradition and ours. First, we are not interested in an unbiased sample of webpages, but in a sample of pages that, taken together, can give a reasonably unbiased picture of a language, independently of whether they are actually 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, we would 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. Second, while here we analyze corpora built via random queries to a search engine, the focus of the paper is not on this specific approach to Web corpus construction, but on the procedure we develop in order to evaluate how varied the linguistic sample we retrieve is. Indeed, in future research it would be interesting to apply our method to corpora constructed using random walks of the Web, along the lines of Henzinger, Bar-Yossef and their colleagues.

3 Measuring distributional properties of biased and unbiased collections

Our goal is to create a "balanced" corpus of webpages from the portion of the Web which contains documents of a given language; e.g., the portion composed of all Italian webpages. As we observed in the previous section, 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 study how the choice of the biased sampling method affects the performance of our procedure. In this section, we introduce our approach by discussing experiments conducted on the BNC where the corpus is seen as a model for the Web, that is, a large collection of documents of different nature. We investigate the distributional properties of the BNC, and the known categories defined within the corpus, which are fully accessible and therefore suitable for random sampling. The method we present highlights important properties that characterize the overall distribution of documents inferrable from incomplete and noisy sampled portions of it; e.g., those which can be retrieved using a suitable set of seed words. In later sections we will show how the method works when the full corpus, the Web, is not available and there is no alternative to "noisy" sampling.

3.1 Collections of documents as unigram distributions

A compact way of representing a collection of documents is by means of a frequency list, where each word is associated with the number of times it occurred in the collection. This representation defines a simple “language model”, a stochastic approximation to the language used in 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. Shannon’s classic example on the entropy of English – Shannon 1948). In this paper we focus on the unigram model as a natural starting point; however the methods we investigate extend naturally to more complex language models.

3.2 Similarity measures for document collections

Our method works by measuring the similarity of collections of documents, approximated as the similarity of the derived unigram distributions, based on the assumption that two similar document collections will determine similar language models. We experimented with two similarity measures over unigram models. The first is the *relative entropy*, or *Kullback Leibler distance* (also referred to as KL), $D(p||q)$ (cf. Cover and Thomas 1991, p. 18), defined over two probability mass functions $p(x)$ and $q(x)$:

$$D(p||q) = \sum_{x \in W} p(x) \log \frac{p(x)}{q(x)} \qquad (1)$$

The relative entropy 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. KL is finite only if the support set of q is contained in the support set of p, hence we make the assumption that the random variables always range over the dictionary W, the set of all word types occurring in the BNC. To avoid infinite cases

Word	Unigram		Total
	P	Q	
w_1	33	17	50
w_2	237	156	393
..	..	..	..
$w_{\|W\|}$	26	1	27
Total	138,574	86,783	225,357

Table 1. Sample contingency table for two unigram distributions P and Q

a smoothing value α is added when estimating probabilities; i.e.,

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

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

Another way of assessing the similarity of unigram distributions is by analogy with categorical data analysis in statistics, where the goal is to assess the degree of *dependency*, or contingency, between two classification criteria. Given two distributions P and Q we create a contingency table in which each row represents a word in W, and each column represents, respectively, frequencies in P and Q (see table 1). If the two distributions are independent from each other, a cell probability will equal the product of its respective row and column probabilities; e.g., the probability that w_1 will occur in distribution P is $p(w_1) \times p(\mathrm{P}) = \frac{50}{225{,}357} \times \frac{138{,}574}{225{,}357} = 0.000135$. The expected number of times w_1 occur in P, under the null hypothesis that P and Q are independent, is then $e_{1,P} = N \times p(w_1)p(\mathrm{P}) = (225,357) \times (0.000135) = 30.48$, as in a multinomial experiment. If the hypothesis of independence is true then the observed cell counts should not deviate greatly from the expected counts. Here we use the X^2 (chi-square) test statistic, involving the $|W|$ deviations, to measure the degree of dependence between P and Q, and thus – intuitively, their similarity:

$$X^2 = \sum_{i,j} \frac{[o_{i,j} - e_{i,j}]^2}{e_{i,j}} \tag{3}$$

Rayson and Garside (2000) use a similar approach to corpus comparison, where deviations in the use of individual words are compared. Here we compare distributions over the whole dictionary to measure the similarity of two text collections.

3.3 Similarity of BNC partitions

In this section we introduce and test the general method in a setting where we can randomly sample from the whole BNC corpus (a classic example of a "balanced" corpus, Aston and Burnard 1998) and from its labeled subsets. The BNC contains 4,054 documents composed of 772,137 different types of words with an overall frequency, according to our tokenization, of 112,181,021 word tokens. Documents come classified along different dimensions. In particular, 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 the purposes of the statistics reported below, we filter out words belonging to a stop list containing 1,430 types and composed mostly of function words. These were extracted in two ways: they either were already labeled with one of the function word tags in the BNC (such as "article" or "coordinating conjunction") or they occurred more than 50,000 times.

Relative entropy and chi-square intuitively measure how similar two distributions are. A simple experiment illustrates the kind of outcomes they produce. If the similarity between pairs of unigrams, corresponding to specific BNC genres or domains is measured, often the results match our intuitions. For example, in

S_meeting				
S_meeting	S_meeting			
R	Genre	KL	Genre	X^2
1	S_meeting	0	S_meeting	0
2	S_brdcast_discuss	0.27	S_interview	82,249
3	S_speech_unscript	0.39	S_parliament	97,776
4	S_unclassified	0.41	S_brdcast_doc	100,566
5	S_interview_hist	0.44	S_speech_unscript	103,843
..	..	..	..	
67	S_demonstration	1.45	W_ac_soc_science	914,666
68	W_fict_drama	1.48	W_pop_lore	973,534
69	S_lect_nat_sci	1.54	W_non_ac_pol_law	976,794
70	S_lect_commerce	1.61	W_misc	1,036,780
71	W_fict_pros	1.64	W_fict_prose	1,640,670

Table 2. Similarities and differences among genres

the case of the genre “S_meeting”[1] the 5 closest (and least close) genres are those listed in table 2.

The table shows that both measures rank higher genres which refer to speech transcriptions of situations involving several people speaking (discussions, interviews, parliament reports, etc.), as is the case with the transcriptions relative to the target category “S_meeting”. On the other hand, at the bottom of the ranking, we find written literary texts, or transcriptions of prepared speeches, which are more dissimilar to the target genre.

Figure 1 plots the matrices of distances between unigrams corresponding to different BNC domains for both X^2 and KL; domains are ordered alphabetically on both x and y axis. Overall the two plots have a somewhat similar topology, resembling a double plateau with peaks on the background. The plot shows, not too surprisingly, that speech transcriptions (whose domain names are

[1] “S_” is the prefix for spoken categories, while “W_” is the prefix for written categories.

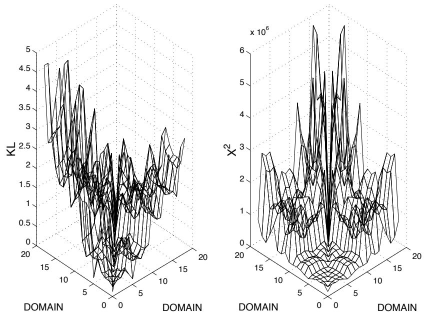

Figure 1. Plots of KL and X^2 distance matrices for the "domain" BNC partitions

prefixed with an "S") tend to be more similar to each other than to written text ("W"-prefixed domains), and vice-versa. However, the figure also shows several important differences between the measures. First of all, X^2 is symmetric while KL is not. In particular, if the size of the two distributions varies greatly, as between the first few domains (close to 1) and the last ones (close to 19) the choice of the background distribution in KL has an effect on the magnitude of the distance: greater if the "true" distribution is larger because of the log-likelihood ratio.

More important is the difference emerging from the region far in the background. Here the two measures give very different rankings. In particular, X^2 tends to interleave the rankings of written and spoken categories. X^2 also ranks lowest several written domains. Table 3 illustrates this fact with an example, where the target domain is "W_world_affairs". Interestingly, X^2 ranks low domains such as "W_commerce" (in the middle of the rank) which are likely to be similar to some extent to the target domain. KL instead produces a more consistent ranking, where all the spoken domains are lower than the written ones and intuitively similar domains such as "W_commerce" and "W_social_science" are ranked highest. One possibility is that the difference is due to the fact that

W_world_affairs				
R	Domain	KL	Domain	X^2
1	W_world_affairs	0	W_world_affairs	0
2	W_soc_science	0.6770	S_demog_unclassified	1,363,840
3	W_commerce	0.7449	S_cg_public_instit	1,568,540
4	W_arts	0.8205	S_cg_education	1,726,960
5	W_leisure	0.8333	W_belief_thought	1,765,690
6	W_belief_thought	1.0405	S_cg_leisure	1,818,110
7	W_app_science	1.0685	S_cg_business	1,882,430
8	W_nat_science	1.4683	S_demog_DE	2,213,530
9	W_imaginative	1.4986	W_commerce	2,566,750
10	S_cg_education	1.5010	W_arts	2,666,730
1	S_cg_public_instit	1.6694	S_demog_C1	2,668,690
12	S_cg_leisure	1.7632	S_demog_C2	2,716,090
13	S_cg_business	1.8945	S_demog_AB	2,834,220
14	S_demog_AB	2.6038	W_soc_science	3,080,840
15	S_demog_C1	2.7667	W_leisure	3,408,090
16	S_demog_C2	2.8110	W_nat_science	3,558,870
17	S_demog_DE	3.2886	W_app_science	3,711,010
18	S_demog_unclassified	4.3921	W_imaginative	5,819,810

Table 3. Rankings produced by KL and X^2 with respect to the domain "W_world_affairs"

the unigram distributions compared with KL are smoothed while raw counts are used for X^2. However, when we tried smoothing the contingency tables for X^2 we obtained even more inconsistent results. An alternative explanation relates the behavior of X^2 to the fact that the distributions being compared have long tails of low frequency counts. It is a matter of contention whether X^2, in the presence of sparse data, i.e., in the presence of cells with less than five counts, produces results which are appropriately approximated by the χ^2 distribution, and thus statistically interpretable (cf. Agresti 1990). It might be that, even if the use described here only aims at relative assessments of dependency/similarity, rather than parametric testing, the presence of large numbers of low frequency counts causes more noisy measurements with X^2 than with KL.

Different metrics have different properties and might provide different advantages and shortcomings depending on the specific task. Since it seems that KL is more appropriate to our task in the remainder of the paper we mainly present results using KL, although we did run all experiments with both measures, often obtaining very similar results.

3.4 A ranking function for sampled unigram distributions

What properties distinguish unigram distributions drawn from the whole BNC from distributions drawn from its subsets – genre, mode and domain? This is an important question because, if identified, such properties might help to discriminate between sampling methods which produce more random collections of documents and more biased ones. We suggest the following hypothesis. Unigrams sampled from the full BNC have distances from biased samples which tend to be lower than the distances of biased samples to other biased samples. If this hypothesis is true then if we sample unigrams from the whole BNC, and from its "biased" subsets, the

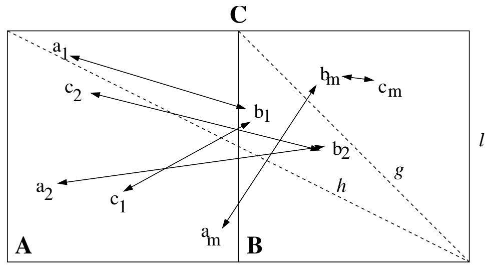

Figure 2. Visualization of the distances (continuous lines with arrows) between points representing unigrams distributions, sampled from "biased" partitions A and B and from the full collection of documents $C = A \cup B$

vector of distances between the BNC sample and all other samples should have lower mean than the vectors for biased samples.

Figure 2 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 random samples of documents 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 lying 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.[2]

[2] Because $h = \sqrt{l^2 + 2l^2} > g = \sqrt{2l^2}$.

Samples from biased portions of the corpus should tend to "remain" in a given region, while samples from the whole corpus should be closer to biased samples, because the unbiased sample draws words from across the whole vocabulary, while biased samples have access to a limited vocabulary. 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 under 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 unbiased sample, while y_j, $j = 2..|Y|$, denote the biased samples. Let $\mathbf{M}$ be a matrix of measurements, $\mathbf{M} \in \mathrm{I\!R}^{|Y| \times |Y|}$, such that $M_{i.j} = \frac{\sum_{k=1}^{N} D(U_{i,k}, U_{j,k})}{N}$, where $D(.,.)$ can be any similarity measure of the kind discussed above, i.e., X^2 or KL. In other words, the matrix contains the average distances between pairs of samples (biased or unbiased). Each row $M_i \in \mathrm{I\!R}^{|Y|}$ contains the average distances between y_i and all other ys, including y_i. We assign a score δ_i to each y_i which is equal to the mean of the vector M_i (excluding $M_{i,j}$, $j = i$):

$$\delta_i = \frac{1}{|Y| - 1} \sum_{j=1, j \neq i}^{|Y|} M_{i,j} \tag{4}$$

It could be argued that also the variance of the distances for y_1 should be lower than the variance of the other ys, because the unbiased sample tends to be equidistant from all other samples. We will show empirically that this seems in fact to be the case. When the variance is used in place of the mean, δ_i is computed as the traditional variance of M_i (excluding $M_{i,j}$, $j = i$):

$$\delta_i = \frac{1}{|Y| - 2} \sum_{j=1, j \neq i}^{|Y|} [M_{i,j} - \mu_i]^2 \tag{5}$$

where μ_i is the mean of M_i, computed as in equation (4).

3.5 Randomness of BNC samples

We first tested our hypothesis on the BNC in the following way. For each of the three main partitions, mode, domain, and genre, we sampled with replacement (from a distribution determined by relative frequency in the relevant set) 1,000 words from the whole BNC and from each of the labels (categories) belonging to the specific partitions. Then we measured the average distance between each label in a partition, plus the sample from the whole BNC. We repeated this experiment 100 times and summarized the results by ranking each label, within each partition type, using δ.

Table 4 summarizes the results of this experiment for all three partitions: mode, domain, and genre (only partial results are shown for genre). The table shows results obtained both with KL and X^2 to illustrate the kinds of problems mentioned above concerning X^2, but we will focus mainly on the results concerning KL. For all three experiments each sample category y_i is ranked according to its score δ_i. The KL-based δ always ranks the unbiased sample "BNC_all" higher than all other categories. At the top of the rankings we also find other less narrowly topic/genre-dependent categories such as "W" (all written texts) for mode, or "W_misc" and "W_pop_lore" for genre. Thus, our hypothesis is supported by these experimental results. Unbiased samples tend to be closer on average to biased samples, and this property can be used to distinguish a biased from an unbiased unigram sampling method. Interestingly, as anticipated in section 3.4, also the variance of the distance vector seems to correlate well with "biased-ness". Unbiased samples tend to have more constant distances from biased samples, than samples to one another. Table 5 summarizes the – comparable – results obtained using for δ_i equation (5); e.g., the variance of M_i.

A different story holds for X^2. There is clearly something wrong in the rankings, although, sometimes we find the unbiased sample ranked the highest. For example, for mode, "S" (spoken) is

Rankings, based on δ-mean						
	Mode		Domain		Genre	
R	X^2	KL	X^2	KL	X^2	KL
1	BNC_all	BNC_all	S_cg_business	BNC_all	S_meeting	BNC_all
2	S	W	S_demog_C1	S_cg_education	S_speech_unscripted	W_misc
3	W	S	S_demog_C2	W_leisure	S_brdcast_discussion	W_pop_lore
4			S_demog_AB	W_arts	S_interview	W_non_ac_soc_sci
5			S_cg_leisure	W_belief_thought	S_unclassified	W_non_ac_humanities_arts
6			S_demog_DE	W_imaginative	S_tutorial	W_newsp_brdsht_nat_misc
7			S_cg_education	S_cg_leisure	S_interview_oral_hist	W_newsp_other_soc
8			S_cg_public_inst	S_cg_business	S_courtroom	W_biography
9			S_demog_unclass	W_app_sci	S_lect_humanities_arts	W_non_ac_nat_sci
10			BNC_all	W_soc_sci	S_brdcast_documentary	W_ac_humanities_arts
11			W_imaginative	S_cg_public_inst	S_lect_soc_sci	W_newsp_other_report
12			no_cat	W_world_affairs	S_parliament	W_newsp_brdsht_nat_arts
13			W_belief	W_commerce	S_brdcast_news	W_newsp_brdsht_nat_soc
14			W_soc_sci	W_nat_sci	S_lect_polit_law_edu	S_brdcast_news
15			W_commerce	S_demog_AB	S_classroom	S_brdcast_discussion
16			W_leisure	S_demog_C1	S_consult	W_newsp_tabloid
17			W_arts	S_demog_C2	S_pub_debate	W_newsp_other_arts
18			W_app_sci	S_demog_DE	S_conv	W_newsp_brdsht_nat_edit
19			W_world_affairs	S_demog_unclass	S_speech_scripted	W_newsp_other_sci
20			W_nat_sci	no_cat	S_sermon	W_newsp_brdsht_nat_report
21					S_demonstration	W_advert
22					W_non_ac_soc_sci	W_ac_soc_sci
23					BNC_all	W_commerce
..					...	...
68					W_fict_drama	S_sportslive
69					W_non_ac_tech_engin	S_consult
70					W_ac_medicine	W_fict_drama
71					W_ac_nat_sci	S_lect_commerce
72					W_fict_poetry	no_cat

Table 4. Rankings based on δ, as the mean distance between samples from the BNC partitions plus samples from the whole BNC; low values for δ ranked higher

Rankings based on δ-variance						
	Mode		Domain		Genre	
R	X^2	KL	X^2	KL	X^2	KL
1	BNC_all	BNC_all	S_cg_public_inst	BNC_all	BNC_all	BNC_all
2	S	W	S_cg_business	W_leisure	W_misc	W_pop_lore
3	W	S	S_cg_education	W_arts	W_non_ac_soc_sci	W_misc
4			BNC_all	W_imaginative	W_non_ac_med	S_brdcast_news
5			S_cg_leisure	W_belief_thought	W_newsp_other_sci	W_non_ac_nat_sci
6			W_belief_thought	S_cg_education	S_brdcast_news	W_non_ac_soc_sci
7			W_imaginative	W_app_sci	W_pop_lore	W_newsp_brdsht_nat_arts
8			W_arts	S_cg_public_inst	W_newsp_brdsht_nat_soc	W_non_ac_humanities_arts
9			no_cat	W_world_affairs	W_newsp_brdsht_nat_sci	W_biography
10			W_leisure	W_soc_sci	S_brdcast_documentary	W_ac_humanities_arts
11			W_soc_sci	W_commerce	W_letters_personal	W_newsp_brdsht_nat_misc
12			W_commerce	W_nat_sci	W_newsp_brdsht_nat_edit	W_newsp_other_soc
13			W_world_affairs	S_cg_business	W_non_ac_humanities_arts	W_essay_school
14			W_app_sci	S_cg_leisure	W_newsp_other_soc	W_fict_prose
15			W_nat_sci	S_demog_unclas	W_biography	W_newsp_brdsht_nat_sci
16			S_demog_unclas	S_demog_AB	W_religion	W_newsp_brdsht_nat_soc
17			S_demog_AB	S_demog_C2	W_essay_school	W_non_ac_med
18			S_demog_C1	S_demog_C1	W_newsp_brdsht_nat_misc	W_fict_poetry
19			S_demog_C2	S_demog_DE	W_non_ac_nat_sci	W_advert
20			S_demog_DE	no_cat	W_essay_univ	W_religion
..					...	...
68					S_interview	S_unclassified
69					S_unclassified	S_lect_commerce
70					S_conv	no_cat
71					S_classroom	S_classroom
72					S_consult	S_consult

Table 5. Rankings based on δ, as the variance of the average distance between samples from the BNC partitions plus samples from the whole BNC; low values for δ ranked higher

ranked higher than “W”, but it seems counterintuitive that samples from only 5% of all documents are on average closer to all samples than samples from 95% of documents. The reason why in general “S” categories tend to be closer (also in the domain and genre experiments) might have to do with low counts as suggested before, and it may also be related to the magnitude of the unigram lists; i.e., distributions made of a small number of unigrams might tend to be closer to other distributions because of the small number of words involved independently of the actual “similarity”.

4 Evaluating the randomness of corpora derived from Google

In our proof-of-concept experiment, we compared the distribution of words drawn from the whole BNC to those of words that belong to various categories. Of course, when we download documents from the Web via a search engine (or sample them in other ways), we cannot choose to sample random documents from the whole Web, nor select documents belonging to a certain category. We can only use specific lexical forms as query terms, and we can only retrieve a fixed maximum number of pages per query. Moreover, while we can be relatively confident that the retrieved pages will contain all the words in the query, we do not know according to which criteria the search engine selects the pages to return among the ones that match the query.[3] All we can do is to try to control the typology of documents returned by using specific query terms (or other means), and we can use a measure such as the one we proposed to look for the least biased retrieved collection among a set of retrieved collections.

[3] If not in very general terms, e.g., it is well known that Google’s PageRank algorithm weights documents by popularity.

4.1 Selection of query terms

Since the query options of a search engine do not give us control over the genre, topic and other textual parameters of the documents to be retrieved, we must try to construct a "balanced" corpus by selecting appropriately balanced query terms, e.g., using random terms extracted from an available balanced corpus (see Sharoff this volume). In order to build specialized domain corpora, we will have to use "biased" query terms from the appropriate domain (see Baroni and Bernardini 2004). We extract the random terms from the clean, balanced, 1M-words Brown corpus of American English (Kučera and Francis 1967). Since the Web is likely to contain much larger portions of American than British English, we felt that queries extracted from the BNC would be overall more biased than American English queries. We extracted the top 200 most frequent words from the Brown ("high frequency" set), 200 random terms with frequency between 100 and 50 inclusive ("medium frequency" set) and 200 random terms with minimum frequency 10 (the "all frequency" set – because of the Zipfian properties of word types, this is a *de facto* low frequency word set). We experimented with each of these lists as ways to retrieve an unbiased set of documents from Google. Notice that there are arguments for each of these selection strategies as plausible ways to get an unbiased sample from the search engine: high frequency words are not linked to any specific domain; medium and low frequency words sampled randomly from a balanced corpus should be spread across a variety of domains and styles.

In order to build biased queries, that should hopefully lead to the retrieval of sets of topically related documents, we randomly extracted lists of 200 words belonging to the following 10 domains from the topic-annotated extension (Magnini and Cavaglia, 2000) of WordNet (Fellbaum, 1998): *administration, commerce, computer science, fashion, gastronomy, geography, military, music, sociology*. These domains were chosen since they look "general" enough to 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.

4.2 Experimental setting

From each source list ("high", "medium" and "all" frequency sets plus the 10 domain-specific lists), we randomly select 20 pairs of words without replacement (i.e., no word among the 40 used to form the pairs is repeated). We use each pair as a query to Google, asking for pages in English only (we use pairs instead of single words to maximize our chances to find documents that contain running text – see discussion in Sharoff this volume). 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 means and variances for the various quantities we calculate.

Our unit of analysis is the corpus constructed by putting together all the non-duplicated documents retrieved with a set of 20 paired word queries. However, the documents retrieved from the Web have to undergo considerable post-processing before being usable as parts of a corpus. In particular, following what is becoming standard practice in Web corpus construction (see, e.g., Fletcher 2004), we discard very large and very small documents (documents larger than 200Kb and smaller than 5Kb, respectively), since they tend to be devoid of linguistic content and, in the case of large documents, can skew the frequency statistics. For technical reasons, we focus on HTML documents, discarding, e.g., PDF files. Moreover, we use a re-implementation of the heuristic used by Aidan Finn's BTE tool[4] to identify and extract stretches of connected prose and discard "boilerplate". In short, the method looks for and selects the fragment of text where the difference between

[4] `http://smi.ucd.ie/hyppia/bte/`

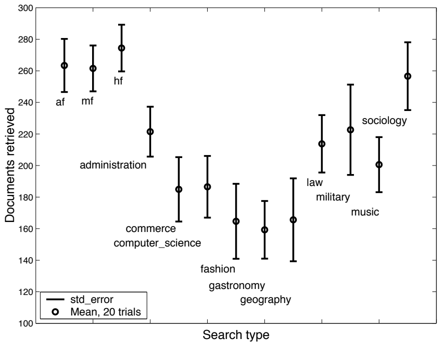

Figure 3. Average number of documents retrieved for each query category over the 20 search sets; the error bar represents the standard deviation

text token count and HTML tag count is maximal. As a further filter, we only keep documents where at least 25% of the tokens in the stretch of text extracted in the previous step are from the list of 200 most frequent Brown corpus words. Because of the Zipfian properties of texts, it is pretty safe to assume that almost any well-formed stretch of English connected prose will satisfy this constraint.

Notice that a corpus can contain maximally 400 documents (20 queries times 20 documents retrieved per query), although typically the documents retrieved are not as many, because different queries retrieve the same documents, or because some query pairs are found in less than 20 documents. Figure 3 plots the means (calculated across the 20 repetitions) of the number of documents retrieved for each query category, and table 6 reports the sizes in types and tokens of the resulting corpora. Queries for the "unbiased" seeds (af, mf, and hf) tend to retrieve more documents, although 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. The difference in number of doc-

Search category	Avg types	Avg tokens
af	35,988	441,516
mf	32,828	385,375
hf	39,397	477,234
administration	39,885	545,128
commerce	38,904	464,589
computer_science	25,842	311,503
fashion	44,592	533,729
gastronomy	36,535	421,705
geography	42,715	498,029
law	49,207	745,434
military	47,100	667,881
music	45,514	558,725
sociology	56,095	959,745

Table 6. Average number of types and tokens in corpora constructed via Google queries

uments retrieved does not seem to have any systematic effect on the resulting distances, as will be briefly discussed in 4.5 below.

4.3 Distance matrices and bootstrap error estimation

We now rank each individual query category y_i, biased and unbiased, using δ_i, as we did before using the BNC partitions (cf. section 3.5). Unigram distributions resulting from different search strategies are 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.

The data collected can be seen as a dataset $\mathcal{D}$ of $n = 20$ datapoints 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. Based on such matrix we can rank a search strategy y_i using δ_i as explained above (cf. section 3.4). Instead of using all n data-points once, we create B "bootstrap" datasets (cf. Duda et al. 2001) by randomly selecting n data-points from $\mathcal{D}$ with replacement (we used a value of B=100). The B bootstrap datasets are treated as independent sets and they are 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 is the mean of the B estimates on the individual datasets:

$$\hat{\delta}_i = \frac{1}{B} \sum_{b=1}^{B} \hat{\delta}_{i,b} \tag{6}$$

Bootstrap estimation can be used to estimate the variance of our measurements of δ_i, and thus the standard error:[5]

$$\sigma_{boot}[\hat{\delta}_i] = \sqrt{\frac{1}{B} \sum_{b=1}^{B} [\hat{\delta}_i - \hat{\delta}_{i,b}]^2} \tag{7}$$

As before we smooth the word counts when using KL, 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] If the statistic δ is the mean, then in the limit of B the bootstrap estimate of the variance is the variance of δ.

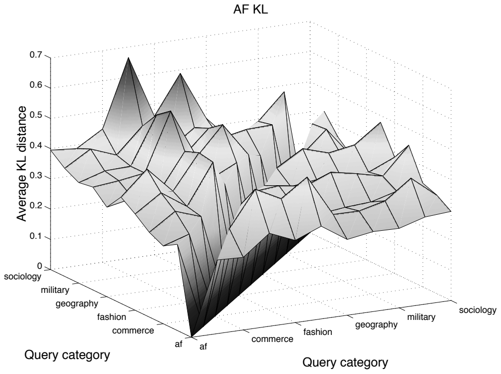

Figure 4. 3D plot of the KL distance matrix comprised of the unbiased query (af) and the biased queries results; only a subset of the biased query labels are shown

4.4 Results

As an example of the kind of results we obtain, figure 4 plots the matrix produced by comparing the frequency lists from all 10 biased queries and the query based on the "all frequency" (af) term set with KL. As expected the diagonal of the matrix contains all zeros, while the matrix is not symmetric. The important thing to notice is the difference between the vectors regarding the unbiased query; i.e., $M_{1,j}$ and $M_{i,1}$ and the other vectors. The unbiased vectors are characterized by smaller distances than the other vectors. They also have a "flatter", or more uniform, shape. The experiments involving the other unbiased query types, "medium frequency" and "high frequency", produce similar results.

The upper half of table 7 summarizes the results of the experiments with Google, compiled by using the mean KL distance. The unbiased sample (af, mf, and hf) is always ranked higher than all biased samples. Notice that the bootstrapped error estimate shows that the unbiased sample is significantly more random than the others. Interestingly, as the lower half of table 7 shows, somewhat similar results are obtained using the variance of the vectors

Rankings with Bootstrap error estimation, $\hat{\delta}$ = mean distance									
R	Sample	$\hat{\delta}_i$	$\sigma_{boot}[\hat{\delta}_i]$	Sample	$\hat{\delta}_i$	$\sigma_{boot}[\hat{\delta}_i]$	Sample	$\hat{\delta}_i$	$\sigma_{boot}[\hat{\delta}_i]$
1	af	0.13040	0.001892	mf	0.12470	0.002176	hf	0.13082	0.002368
2	commerce	0.14997	0.007186	commerce	0.15062	0.007273	commerce	0.14989	0.007177
3	geography	0.16859	0.009102	geography	0.16986	0.009061	geography	0.16907	0.009152
4	admin	0.17254	0.004040	admin	0.17338	0.004081	admin	0.17257	0.004035
5	fashion	0.17292	0.007944	fashion	0.17403	0.007981	fashion	0.17313	0.007893
6	comp_sci	0.17437	0.004554	comp_sci	0.17486	0.004651	comp_sci	0.17408	0.004605
7	military	0.19181	0.007113	military	0.19388	0.007160	military	0.19192	0.006976
8	gastronomy	0.19560	0.009307	gastronomy	0.19708	0.009461	music	0.19612	0.006689
9	music	0.19583	0.006611	music	0.19761	0.006754	law	0.19635	0.005669
10	law	0.19707	0.005661	law	0.19900	0.005718	gastronomy	0.19646	0.009335
11	sociology	0.24075	0.008674	sociology	0.24329	0.008596	sociology	0.24023	0.008496
Rankings with Bootstrap error estimation, $\hat{\delta}$ = variance									
R	Sample	$\hat{\delta}_i$	$\sigma_{boot}[\hat{\delta}_i]$	Sample	$\hat{\delta}_i$	$\sigma_{boot}[\hat{\delta}_i]$	Sample	$\hat{\delta}_i$	$\sigma_{boot}[\hat{\delta}_i]$
1	af	0.00018	3.06478e-05	mf	0.00023	3.03428e-05	hf	0.00019	3.07505e-05
2	music	0.00028	2.52644e-05	music	0.00026	2.26974e-05	music	0.00028	2.47110e-05
3	commerce	0.00029	4.45361e-05	commerce	0.00031	4.10216e-05	commerce	0.00030	4.39917e-05
4	fashion	0.00043	6.83792e-05	fashion	0.00043	6.94822e-05	fashion	0.00043	7.07811e-05
5	geography	0.00046	6.43744e-05	geography	0.00046	6.61234e-05	geography	0.00044	6.74963e-05
6	gastronomy	0.00066	7.31346e-05	gastronomy	0.00065	6.90454e-05	gastronomy	0.00062	6.39204e-05
7	comp_sci	0.00068	5.57851e-05	comp_sci	0.00075	6.14489e-05	comp_sci	0.00072	5.73490e-05
8	admin	0.00079	8.58979e-05	admin	0.00082	8.32991e-05	admin	0.00081	8.86421e-05
9	military	0.00094	0.000114039	military	0.00091	0.000116997	military	0.00095	0.000120596
10	law	0.00147	0.000145864	law	0.00145	0.000152849	law	0.00154	0.000152587
11	sociology	0.00296	0.000295807	sociology	0.00293	0.000307310	sociology	0.00302	0.000323409

Table 7. Google experiments: rankings for each unbiased sample category with bootstrap error estimation (B=100)

M_i instead of the mean, to compute δ_i. The unbiased method is always ranked highest. However, since the specific rankings produced by mean and variance show some degree of disagreement, it is possible that a more accurate measure could be obtained by combining the two measures.

4.5 Discussion

We observed, on Google, the same behavior that we saw in the BNC experiments, where we could directly sample from the whole unbiased collection and from biased subsets of it (documents partitioned by mode, domain and genre). 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.

If our measure is quantifying unbiased-ness, then the lower the value of δ with respect to a fixed set of biased samples, the better the corresponding seed set should be for the purposes of unbiased corpus construction. In this perspective, our experiments also show that unbiased queries derived from "medium frequency" terms (e.g., *places*, *wonderful*) perform better than all frequency (therefore mostly low frequency) and high frequency terms (e.g., *soils*, *contraction* and *even*, *what*, respectively). Thus, while more testing is needed, our data provide some support for the choice of words that are neither too frequent nor too rare as seeds, when building a Web-derived corpus.

Finally, the results indicate that, despite the fact that different query sets retrieve on average different amounts of documents, and lead to the construction of corpora of different lengths, there is no sign that these differences are affecting our δ measure 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 (the unbiased samples) and at the bottom of the ranks (law, sociology) in table 7.

5 Conclusion

As research based on the Web as corpus, and in particular on automated Web-based corpus construction, becomes more prominent within computational and corpus-based linguistics, many fundamental issues have to be tackled in a more systematic way. Among these, there is the problem of assessing the quality and nature of a corpus built with automated means.

In this paper, we considered one particular approach to automated corpus construction (via search engine queries for combinations of a set of seed words), and 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 frequency distribution of words in an unbiased collection will be, on average, less distant from distributions derived from biased partitions, than any of the biased distributions (we showed that this is indeed the case for a collection where we have access to the full unbiased and biased distributions, 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, besides confirming the hypothesis that corpora created using unbiased seeds have lower average distance to corpora created using biased seeds, compared to the average distance of each biased corpus to the others biased corpora, suggest that the seeds to build an unbiased corpus should be selected among medium frequency words (medium frequency in an existing balanced corpus, that is), rather than among high frequency words or words not weighted by frequency (as in the setting in which we sampled from the whole Brown type list).

We realize that our study leaves many questions open, each of them corresponding to an avenue for further study. One of the crucial issues is 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 is unlikely to be balanced in terms of topics and genres. Issues of representativeness and balance of corpora are widely discussed by corpus linguists (see Kilgarriff and Grefenstette 2003 for an interesting perspective on these issues from the point of view of Web-based corpus work). 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 of unbiased-ness/balance is appropriate, all it tells us is that a certain corpus is more/less biased than another corpus with respect to the biased corpora we compared them against (e.g., in our case, the corpus built with mid frequency seeds is less biased than the others with respect to corpora that represent 10 broad topic-based WordNet categories). Thus, it will be important to check whether our methodology is stable across choices of biased samples. In order to verify this, we plan to replicate our experiments using a much higher number of biased categories, and systematically varying the biased categories. We believe that this should be made possible by sampling biased documents from the long lists of pre-categorized pages in the Open Directory Project (`http://dmoz.org/`).

Our WordNet-based queries are obviously aimed at creating corpora that are biased in terms of *topics*, rather than *genres* or *textual types*. A balanced corpus should also be unbiased in terms of genres. In order to apply our method to genre-based balancing, we need to devise ways of constructing corpora that are genre-specific, rather than topic-specific. This is a more difficult task, not least because the whole notion of what exactly is a “Web genre” is far from settled (see, e.g., Santini 2005). Moreover, while sets of seed words can be used to retrieve words belonging to a certain topic, it is less clear how genres can be targeted through search engine queries. Again, the Open Directory Project categorization could be helpful here, as it seems to be, at least in part, genre-based (e.g., the Science section is organized by topic – agriculture, biology, etc. – but also into categories that are likely to correlate, at least partially, with textual types: chats and forums, educational resources, news and media, etc.)

We tested our method on three rather similar ways to select unbiased seeds (all based on the extraction of words from an existing balanced corpus). Corpora created with seeds of different kinds (e.g., basic vocabulary lists, as in Ueyama this volume) should also be evaluated. Indeed, a long term goal would be to use our method to iteratively bootstrap "optimal" seeds, starting from an arbitrary seed set. More in general, the method is not limited to the evaluation of corpora built via search engine queries. For example, it would be interesting to compare the randomness of corpora built in this way to that of corpora built by Web crawls that start from a set of seed URLs (e.g., Emerson and O'Neil this volume).

Finally, we would like to explore extensions of our method that could be applied to the analysis of corpora in general (Web-derived or not), both for the purpose of evaluating their relative degree of biased-ness, and as a general-purpose corpus comparison technique (on corpus comparison, see, e.g., Kilgarriff (2001).

References

Agresti, A. (1990). *Categorical data analysis*, New York: Wiley.

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

Bar-Yossef, Z., Berg, A., Chien, S., Fakcharoenphol, J. and Weitz,

D. (2000). Approximating aggregate queries about Web pages via random walks. *Proceedings of VLDB 2000*, 535-544.

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

Bharat, K. and Broder, A. (1998). A technique for measuring the relative size and overlap of the public Web search engines. *Proceedings of WWW7*, 379-388.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8, 1-15.

Cover, T.M. and Thomas, J.A. (1991). *Elements of information theory*, New York: Wiley.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001). *Pattern classification, 2nd ed.*, New York: Wiley.

Fellbaum, C. (ed.) (1998). *WordNet: An electronic lexical database*, Cambridge: MIT Press.

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

Ghani, R., Jones, R. and Mladenić, D. (2001). Mining the Web to create minority language corpora. *Proceedings of the 10th International Conference on Information and Knowledge Management*, 279-286.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics* 6, 1-37.

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

Henzinger, M., Heydon, A. and Najork, M. (2000). On near-uniform URL sampling. *Proceedings of WWW9*, 295-308.

Kučera, H. and Francis, W.N. (1967). *Computational analysis of present-day American English*, Providence: Brown University Press.

Lee, D. (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.

Magnini, B. and Cavaglia, G. (2000). Integrating subject field codes into WordNet. *Proceedings of LREC 2000*, 1413-1418.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, 1-6.

Santini, M. (2005). Genres in formation? An exploratory study of Web pages using cluster analysis. *Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27.

Sharoff, S. (Submitted). Open-source corpora: Using the Net to fish for linguistic data.

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