Min(d)ing English language data on the Web: What can Google tell us?

Gunnar Bergh
Mid-Sweden University

1 Introduction

As commonly recognized, the era of modern corpus linguistics is approaching the half-century mark. During the past 50 years, we have witnessed a series of important landmark events in this field, ranging from the early attempts at mechanolinguistics by Juiland and Busa in the 1950s, to the pioneering work on computerized corpora in the 1960s and 1970s, involving first written material in terms of the *Brown Corpus* and the *LOB Corpus*, and later spoken material in connection with the *London-Lund Corpus*; in the 1980s, we have experienced the large-scale corpus projects of *Cobuild* and the *Bank of English*, and in the 1990s the *British National Corpus* (BNC) and the *International Corpus of English* (ICE) (e.g. McEnery and Wilson 2001: 20 ff.).

Having now entered the 21st century, it is clear that there are new challenges ahead for the corpus linguist. In terms of standard corpora, for example, we know that the *American National Corpus* (ANC) is under development, a parallel to the BNC with 100 million words of transatlantic English (e.g. Ide et al. 2002), and there is also a great deal of work going on with sophisticated varieties of learner corpora and multilingual (parallel) corpora (e.g. Botley et al. 2000; Granger 2004). However, the biggest challenge of today is undoubtedly the growing body of text-based information available on the World Wide Web (henceforth the Web). While originally intended as a pure information source only, this material forms in fact the largest store of textual data in existence, and as such it constitutes a tantalizing resource for various linguistic purposes.

Let us look at some initial figures. As regards the size of the material on the Web, a rough estimate indicates that there are currently (December 2004) about eight billion Web pages available (cf. http://www.google.se/press/funfacts.html), containing perhaps as much as 50 terabytes of text: at a generous average of 10 bytes per word (cf. Kilgarriff and Grefenstette 2003), these figures suggest a body of no less than five trillion (5 000 billion) words in one form or another.

Out of this massive multilingual collection of texts and text fragments, it appears that about two thirds are written in English (e.g. Xu 2000), although the proportion of non-English texts seems to have increased in recent years (e.g. Grefenstette and Nioche 2000). This means that there is probably something in the range of 3 000 billion words of English to be found on the Web, forming a virtual English supercorpus ready for use by enterprising linguists in all manner of language research (cf. Bergh et al. 1998).

But what, then, can we use the Web material for more specifically? Judging from recent work in this field, there are two types of linguistic usage that seem particularly profitable. On the one hand, it can be used as a pure textual resource, providing raw material for different types of corpus building, either in terms of domain-specific corpora (e.g. Kilgarriff 2001) or corpora representing the language in some general sense (e.g. Fletcher 2004). On the other hand, and equally important, it can be used for investigation of various aspects of current language usage, notably in terms of frequency-based patterns: one case in point is the study of rare or neologistic language, i.e. elements and structures which are either very infrequent (e.g. Bergh et al. 1998) or have been very recently coined (e.g. Renouf 2003).

One problem in this context, however, is that the Web turns out to be a somewhat intractable collection of textual material, as witnessed by those corpus linguists who have tried to access it through available search engines. This is mainly due to the fact that it constitutes a rather haphazard accumulation of digital text. Put more specifically, it consists of a heterogeneous, non-sampled body of text-based information in a variety of different formats which is multilingual, contains lots of duplicates or near-duplicates, and which is continuously changed and updated (e.g. Kilgarriff 2001, Renouf 2003, Fletcher 2004). Yet, as already noted, it is by far the world's largest store of texts, "the ultimate monitor corpus" in Sinclair's (1991) terminology, freely available and maximally broad in topicality and domain coverage.

Now, let us say that we want to focus on language structure as such. We may then ask ourselves why we need to consider the Web at all, as there are already comprehensive standard corpora at hand featuring millions of words, for example the Cobuild corpus and the BNC? Let us take a simple example to illustrate the problem. If we are interested in recent loan words in English, say, the word *Taliban*, and in particular its concord properties, it is a simple matter to produce a relevant search string and look it up in either of these corpora. The only problem is that there are rather few results returned. This is to say that a search for *Taliban* in the 56 million words of the Cobuild corpus yields only some 40 matches in all, of which only four provide clear-cut information on concord

issues (three indicating plural concord and one singular concord), a figure which is clearly too low for reliable conclusions to be drawn (e.g. Kilgarriff and Grefenstette 2003). And turning to the BNC, with its 100 million words of text, we are faced with an even more disappointing outcome – there are no matches at all to be found. Yet, we know that the word *Taliban* has been in frequent use lately, in particular in connection with the 9/11 events (e.g. Ohlander and Bergh 2004), thus highlighting the problem with many corpora of today, i.e. that they become dated rather quickly, not least with regard to the representation of current vocabulary.

What can we do in this situation? The natural remedy is, of course, to turn to a significantly larger collection of texts. And, as we already know, this is exactly what the Web provides – with emphasis on significantly! With the previous figures of size in mind, and restricting ourselves to English, it is clear that the Web is tens of thousands of times bigger than the most comprehensive corpora of today, and that it contains a multitude of material from much more recent dates. What happens, then, if we look up the word *Taliban* in this enormous Web-based material? In fact, no less than 1 890 000 matches from English Web pages are returned (2 080 000 matches if other languages are allowed), illustrating a range of different types of patterns and constructions where this word occurs (December 2004). This is a truly amazing figure in relation to what standard corpora are able to return, thus confirming the size and freshness of the Web. We will return to the word *Taliban* later.

With the availability and potential usefulness of the Web established, we need to consider next the most fruitful ways of accessing it. This brings us to the problem of available search software, and what services this software is able to provide. Whereas standard corpora typically come with software specialized for searches for different linguistic forms, search engines on the Web are designed to find contents, using the linguistic form only as a means to achieve that goal. Thus, we know that the typical corpus search software, be it MonoConc, WordSmith or SARA, is equipped with algorithms to present its output in the form of KWIC concordances, whereas search engines, such as AltaVista, WebCrawler or FAST, are built primarily for information retrieval, presenting their results in the form of lists of more or less context-free hyperlinks only. From the point of view of corpus linguistics, such hyperlink lists are awkward, since they often necessitate tedious individual look-up of target words and constructions, which is a very time-consuming process. In recent years, however, this problem has been partly alleviated by the emergence of more sophisticated search technology which is designed to supply proper context to target strings, thus emulating the user-friendliness of traditional concordancing programmes.

These observations bring us to the main aim of the present paper, i.e. to take a glimpse at the most recent applications for using the Web as a source for linguistic work. In our case, this means looking at Google, arguably the most potent search engine available at present, together with a recently developed Web concordancing system, WebCorp, which is able to run on top of it.

2 Description of Google

Let us start by saying a few words about the development of the Google search engine as a miner of language data on the Web. Questions of particular interest in this context include: How does the search engine work? What does it find (and what is left out)? How are the results ranked and displayed? To what extent are specialized searches possible?

Founded at Stanford university in the late 1990s by Sergey Brin and Larry Page, Google was created mainly in order to scale with the dramatic growth of the Web, and to keep up with the increasing number of information search queries from the general public. To illustrate this growing demand, we may note that in 1994 a search engine like the World Wide Web Worm (WWWW) received about 1 500 queries per day; in 1997 the AltaVista engine received about 20 million queries per day (cf. Brin and Page 1998), a figure which is to be compared to the hundreds of millions of queries handled by Google today. Another purpose of Google was to remedy some of the experienced shortcomings of previous engines, for example that they were insufficient in terms of capacity (as regards both speed and size), that they returned too many low-quality hits (or "junk results"), which were also difficult to take in at first glance, and that they were too easily manipulated by advertisers.

In view of these problems, Brin and Page launched a new type of server setup that was capable of handling the extremely large data sets on the Web in more efficient manner. By using a large network of interlinked PCs, they implemented an automated crawler-based engine that was specifically designed for rapid indexing, search responsiveness and relevance ranking of results. The outcome of this work is shown in Figure 1, which gives a sketch of the basic anatomy of Google:

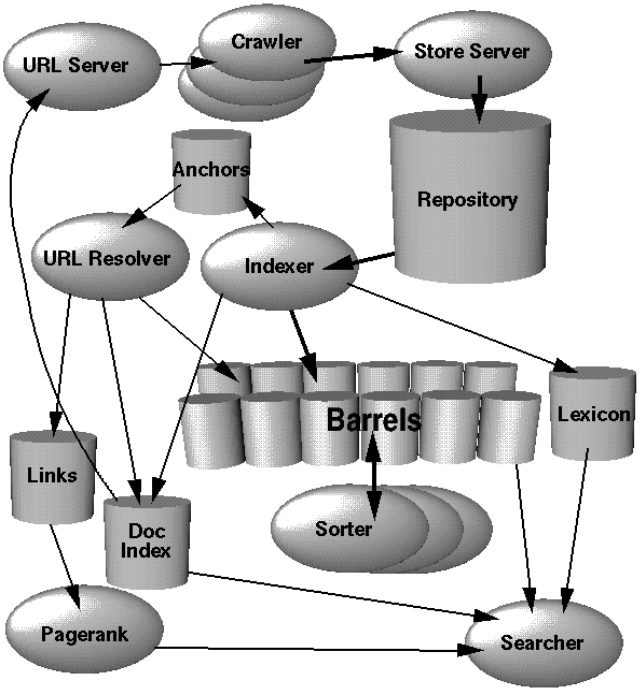

Figure 1: Diagramme of Google's fundamental architecture as presented in Brin and Page (1998)

The spider, or robot, of the system, often referred to as GoogleBot, consists of a set of fast crawlers which are able to follow hyperlinks to Web pages, read those pages, and download their contents. The process is initiated by an URL server, which provides lists of URLs to be explored by the crawlers. When pages have been fetched and provided with a unique tag (a docID), they are sent to a store server, which compresses the pages and stocks them in a repository. The indexing function is carried out by a multifunctional indexer, which is designed to read the repository, uncompress the documents and parse them into a set of "word occurrences", or hits, where information is recorded on document position, font size, capitalization, etc. The indexer then accumulates these hits in a number of barrels, thereby composing a preliminary and partially sorted forward index. At the same time it parses out all the links in every document and saves

the resulting link structure (i.e. where each link is pointing to and from) in an anchors file. A URL resolver reads the anchors file, and via different subprocesses it produces a links database which forms the basis of the relevance ranking of the different Web pages (so-called PageRank). Finally, a sorter is brought in to produce an inverted index from the forward index in the barrels. Generating a set of wordIDs in the new index, the sorter feeds this information into a program which compiles a lexicon to be used by the searcher function. It is this searcher, in collaboration with PageRank, which is employed by Web servers when search queries from the general public are answered by Google (Brin and Page 1998).

From this functional sketch, it is obvious, then, that Google features an advanced system for organising and giving access to the huge amounts of information available on the Web. As already noted, it is estimated that the engine currently maintains an index of eight billion Web pages, a figure which can be compared to the 110 000-page index of the WWWW in 1994 and the two-million-page index of WebCrawler in 1997 (cf. Brin and Page 1998). At peak speed, it is able to crawl hundreds of Web pages per second, keeping a very large number of connections open at the same time. This high capacity has clearly benefited the user in a direct sense, since it enables the system to return comprehensive lists of results within one or two seconds, sometimes even within a split second. Incidentally, these performance data also explain why the system is called *Google*: reflecting the idea of a very large-scale search engine, the name is simply a spelling variant of the word *googol*, meaning 'the number 1 followed by 100 zeros or, in other words, 10 to the power of 100'.

The next question we need to ask ourselves is what Google is able to find. Obviously, this is dependent on the type of search string keyed in by the user. Here we find the standard possibilities of many corpus search programmes, i.e. that we can make one or more words form a search string. If there is more than one word in the string, it is possible to distinguish between approximate searches, i.e. where the search words do not need to be adjacent to each other, and exact ones, i.e. where they are necessarily adjacent to each other, the latter variant then requiring a search string within quotation marks. If we key in, say, the phrase *corpus linguistics*, the following list of results is returned, shown here in terms of the first five hits (January 2004):

Figure 2: Top-five list of Web pages returned by Google on the basis of the search term "corpus linguistics"

As we can see, Google returns about 18 000 matches here, embedded in a short stretch of context. This figure can be compared to the about 11 000 matches from a corresponding search by the competing AltaVista engine, showing the significantly greater index coverage of Google. The output also has a higher level of precision, generally speaking, in that it is restricted to showing only those results which feature all the search terms keyed in: unlike the AltaVista, there are no near-hits, i.e. cases where only one of the search terms is represented. Clearly, this spares the user the frustration of viewing a multitude of

partly irrelevant hits. A further advantage is that Google does not sell placement within results to advertisers: as a counter-illustration of this stance, we can use to the AltaVista output referred to above, whose result list for "corpus linguistics" is in fact topped by two instances of advertising material. It goes without saying that this is not an ideal situation for corpus linguists, who generally prefer a more objective approach to empirical investigations. On the down side of the engine, we may note that Google is not case-sensitive and does not offer wildcards or stemming strategies as a way of customising search strings; i.e. it is not possible to use strings such as *Googl** or *Goog* to find variants or derivatives of the word *Google*. Also, referring to its previously outlined functionality, it is clear that the engine can only retrieve search terms which have been stored in its periodically updated index, which means that very recent additions to the Web may not be included in the output (cf. the case of *Sophiegate* in Renouf 2003).

With the formal range of results settled, we may proceed to look into the question of how these data are presented to the user. As some searches are likely to yield thousands of relevant hits, it is necessary to establish some kind of principle for the ordering of them. While corpus software proper usually yields objective results from corpus searches – either in terms of all the hits or a random sample of them, displayed either alphabetically or on the basis of their place of occurrence in different texts – Google relies on an intricate system of voting for this function, what has already been referred to as PageRank (cf. above). To illustrate how this works, let us return to the example with the phrase *corpus linguistics* and its top-five hit list given in Figure 2. Put plainly, why is Michael Barlow's site rated higher here than that reflecting the work of Tony McEnery and Andrew Wilson? Arguably, this is due to the ranking system. In simplified terms, PageRank relies on the so-called democratic nature of the Web by using its enormous link structure as a gauge of the value of an individual page. Basically, Google interprets a link from one page to another as a vote, by the first page, for the second page. But the sheer volume of votes, or links, a page receives is not the only measure used; the system also explores the page that casts the vote. The purport of this is that votes cast by pages that are themselves popular weigh more heavily and help to make other pages "important". Popular sites thus receive a higher page rank, something which Google remembers each time it performs a search. However, since such pages mean nothing if they do not match the original query, Google combines PageRank with its text-matching techniques to find pages that are both popular and relevant to the search in question (cf. http://www.google.com/technology/index.html). In the case of Barlow contra McEnery and Wilson, then, it seems to be the case that while both pages feature the phrase *corpus linguistics* in headline fashion, the

former is given a more prominent position on the hit list because it is ranked higher, apparently due to there being more links pointing to it from other "important" pages.

All in all, then, we may conclude that PageRank is a rather ingenious system for arranging search results. Yet, it has a somewhat subjective touch to it that is slightly problematic from the point of view of corpus linguistics: as users tend to restrict their viewing of search results to the initial part of the list, whether it is the first ten, 100 or 1 000 hits, it follows that the ranking of Web pages is likely to favour linguistic constructions which happen to occur on more popular pages, thereby risking a certain bias in studies based on language data mined by Google.

Finally, in the description of the capabilities of Google, it may also be worth while to look at its ability to slice the Web in different ways, i.e. to support specialized search initiatives. Such searches typically imply that some sort of restriction has been imposed on a query, in terms of language, file format, URL, date, domain, content, etc. Table 1 shows some of the possibilities offered, illustrated through the results of a set of corresponding restricted searches for the string "corpus linguistics" performed through Google's Advanced Search interface (January 2004):

Table 1: Absolute distribution of the about 18 000 Google matches of the search term "corpus linguistics" as a function of a set of restriction (slicing) parameters

Parameter	Restriction	Search result
Language	English-only (cf. German-only)	15 200 hits (541 hits)
File format	PowerPoint-only (.ppt)	67 hits
Place	Titles-only (cf. URLs-only)	869 hits (164 hits)
Date	Updated in the past three months	7 220 hits
Domain	Australia (.au)	194 hits
Content	SafeSearch excluding "adult" sites	17 200 hits
Similar pages	the Barlow site (cf. the Google site)	29 hits (30 hits)
Linking pages	the Barlow site (cf. the Google site)	840 hits (230 000 hits)

These spot-test figures yield a rough idea of in what circles of cyberspace we can expect to find the phrase *corpus linguistics*. Thus, not surprisingly, the term seems to be frequent in the textual body of English-language Web documents, notably pages which are fairly recent and "safe" (sic!), although it is seldom realized in documents from the Australian domain or in PowerPoint format (probably because such documents are not too frequent themselves in this context). More importantly, however, the existence of this type of slicing possibilities accentuates Google's potential as a versatile tool for various forms of empirical language research, not least in cases of language-specific or domain-specific study.

3 Description of WebCorp

Having looked at the main features and abilities of Google, let us turn next to a profitable development in this field, namely the WebCorp tool, i.e. a Web concordancer which runs on top of Google (and a set of other search engines). Designed by the Research and Development Unit for English Studies (RDUES) at Liverpool, this search tool is intended to provide contextualized examples of language usage from the Web, and to present them in a form tailored for linguistic analysis. As such, then, it adds a layer of refinement to standard Web searches by offering a wider range of search possibilities and presentational means than search engines proper.

The immediate question in this context is of course how this is possible. In order to break into the functional circle of WebCorp, let us start by considering a sketch of its infrastructure (Figure 3):

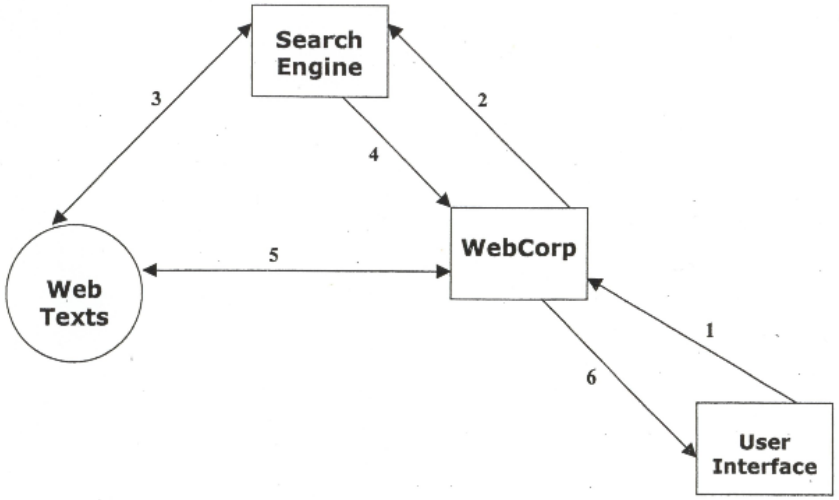

Figure 3: Diagramme of WebCorp's basic architecture as presented in Renouf (2003) and Renouf et al. (2004)

As the graph shows, WebCorp works in a set of different stages. When a query is formulated, the tool first interfaces with the search request and converts it to a format acceptable to search engines. It then "piggy-backs" on the selected search engine, which finds the search term through its index and provides a URL for the relevant source text. WebCorp downloads that text temporarily, extracts the search term and the appropriate linguistic context, collates it, and finally presents it to the user in desired format, prototypically as a KWIC concordance (Renouf 2003; Renouf et al. 2004).

Trying to home in on the finer details of the tool, we may begin by noting that its scope is rather limited. While ordinary search engines are able to process millions of search string matches (although not all of them are presented to the user), WebCorp is limited to treating results from a maximum of 120 Web pages only. A surprisingly low figure though this may seem, the limit has been set at this level for reasons of processing speed, i.e. at present a larger number of pages would simply make the search phase too time-consuming. As a means of illustration, let us key in the search term "corpus linguistics" again and check how the tool organizes the corresponding output list (January 2004), reflected in Figure 4:

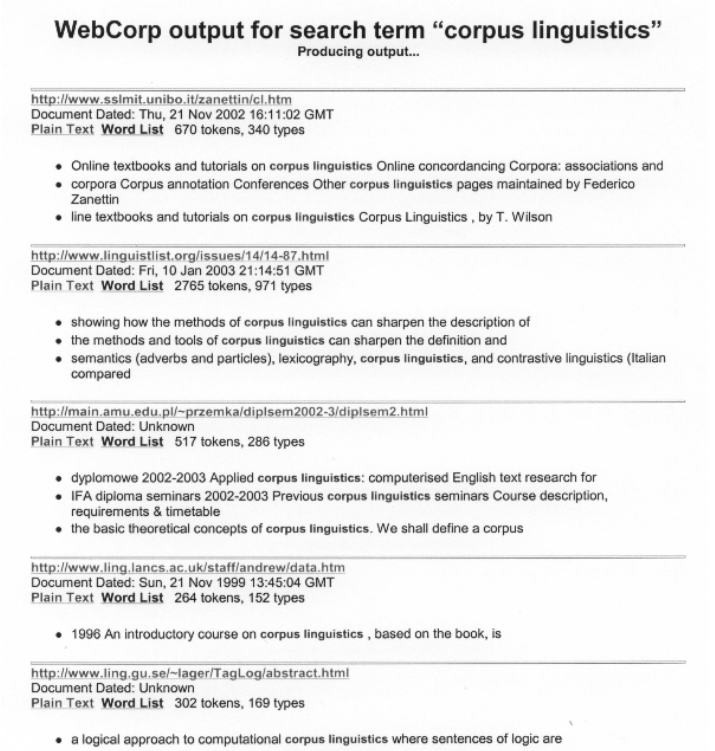

Figure 4: Top-five list of Web pages returned by WebCorp via Google on the basis of the search term "corpus linguistics"

From this sample, it is immediately obvious that WebCorp performs a more thorough analysis of Web pages than corresponding search engines. Apart from attaching useful information in terms of date and type-token ratio, the tool provides the user with a complete KWIC concordance featuring all the search string matches found, listed according to the respective Web pages in which they occur. With reference to the cropped data above, this means that WebCorp found 11 matches in five pages, all of them presented in a pre-defined amount of context. The ordering of the Web pages attracts similar interest. When Google is used as the target engine, the choice of these pages is based on the PageRank system; i.e. the most popular pages according to Google are also the ones that WebCorp will use to create its concordances. This means that whatever bias

Google is responsible for in the ranking of Web pages will automatically transfer to the WebCorp data. The only difference is, as hinted at above, that while Google is able to find only one search string match per page, WebCorp will find all the matches in each of the pages; i.e. on the basis of its maximum of 120 pages, WebCorp might very well find 500 examples of a sought-for item.

Another important property of WebCorp is its ability to produce clear collocational profiles. A standard ingredient in many types of lexical research, such profiles consist of frequency counts of collocates within a pre-determined textual span to a search term. By way of illustration, consider Figure 5, once again based on the search string "corpus linguistics" (January 2004), featuring results in the left-right positions of a +4/-4 span (cf. Renouf et al. 2004).

Word	Total	L4	L3	L2	L1	R1	R2	R3	R4	Left Total	Right Total
Linguistics	55	11	5	9	8	2	5	5	10	33	22
Corpus	47	7	6	3	1	8	6	7	9	17	20
English	42	2	1		33		1	3	2	36	6
University	41	5	2	2	2	5	5	15	5	11	30
linguistics	27	4	4	4	5		2	5	3	17	10
corpus	24	4	3	4		1	4	1	7	11	13
Language	23		1	1	1	1	13	5	1	3	20
Centre	23		9	12				2		21	2
eds	22		1	15	6					22	0
Corpora	21	4		2	1	5	2	5	2	7	14
2001	21	2	3		1	10	2		3	6	15
Studies	20			6	1	12		1		7	13
language	17	3	2	1		1	6	4		6	11
Jan	16	13	2	1						16	0

Altenberg	15	10	3							2	13	2
1998	15	2	2		9				1	1	13	2
McEnery	15	3	2	2			3	2	3	7	8	
Introduction	15	1	1	9			4			11	4	
Directions	15			15						15	0	
course	14	2		3	1	5	1	1	1	6	8	

Figure 5: Top collocates of "corpus linguistics" (excluding stopwords) as calculated by WebCorp via Google

As we can see, the collocational profile of *corpus linguistics* is both useful and thought-provoking. It is useful in the sense that it provides a guide to some of the combinatorial properties of the target noun phrase, e.g. that *English corpus linguistics* and *corpus linguistics course* are recurrent combinations. And it is thought-provoking in the sense that it sets the scene for the methodology concept as such, e.g. that the primary language of corpus linguistics is English, that corpus linguistics is relevant for universities and centres, that the important years of corpus linguistics include 1998 and 2001 (because of the publication of two significant textbooks with that title, one by Biber, Conrad and Reppen (1998) and the other by McEnery and Wilson (2001)), and that influential people in corpus linguistics go by the names of "Altenberg" and "Jan" (the former clearly referring to Bengt Altenberg, and the latter presumably to the two Jans with the surnames Svartvik and Aarts, respectively). And perhaps most significant of all: this type of collocational profiling cannot be generated through Google alone, nor through any other available search engine.

Finally, with regard to the capabilities of WebCorp, we may also take a cursory look at the possibilities of performing specialized searches. One aspect of this is the use of wildcards, which the tool supports; another is the slicing of the Web in terms of restricting searches to particular topics or domains (cf. Renouf et al. 2004). In Table 2, a few relevant examples of the latter type are given.

Table 2: Absolute distribution of WebCorp matches of the search term *corpus linguistics* as a function of a set of restriction parameters

Parameter	Restriction	Search result
Search engine	Google AltaVista	146 hits in 96 pages 66 hits in 50 pages
Site domain	Australia (.au)	78 hits in 34 pages
Newspaper domain	UK broadsheets	2 hits in 2 pages
Textual domain	Science	6 hits in 7 pages

4 The Web versus standard corpora

Keeping in mind how Google and WebCorp are structured and how they deal with results, we may now proceed to the more general issue of the linguistic representativeness of Web data as mined by Google. In other words, the question in focus is whether we can make such data replicate frequency patterns derived from sample-based collections of language, i.e. standard corpora. And if differences occur, how can they be accounted for?

In order to look into these questions, let us take a simple example with the word *colour*, spelt both the American and the British way. The first thing we need to investigate, then, is the relationship between these two spelling variants in some fairly recent corpora. For this purpose, the one-million-word Frown and FLOB corpora were selected, representing American and British English, respectively, as well as the 100-million-word British National Corpus. Table 3 shows the outcome of this simple spot check.

Table 3: Comparison of the frequencies of the two spelling variants *colour* and *color* in three different corpora of English, as mined by the concordancing programmes MonoConc Pro and SARA

	Frown	%	FLOB	%	BNC	%
colour	3	2.3	111	99.1	11 345	99.0
color	125	97.7	1	0.9	115	1.0
Total	128	100	112	100	11 460	100

As we can see, these corpus figures show quite clearly that the major spelling alternative in each case, *color* in American English and *colour* in British English, tends to occur as often as 98 or 99 times per 100 recorded instances of the lexeme. And conversely, it follows that the minor spelling alternative, *colour* in American English and *color* in British English, reveals the opposite pattern, with only one or two occurrences per 100 instances. We may hence assume that these proportions represent the real distribution of these spelling alternatives in the two regional varieties at issue.

Let us next investigate how these figures compare to the results of corresponding searches of the Web, made separately for American and British English through Google's language variety restriction facility (January 2004), illustrated in Table 4:

Table 4: Comparison of the frequencies of the two spelling variants *colour* and *color* in American and British language material published on the Web, as mined by Google

	US English	%	UK English	%
colour	1 880 000	23	1 870 000	86
color	6 240 000	77	303 000	14
Total	8 120 000	100	2 173 000	100

In the 10 million range as regards matches, these statistics yield a slightly deviant picture. Whereas the Google data basically conform to the pattern derived from the sampled corpora, the proportions of the two spelling alternatives are now significantly different, i.e. the major spelling variant in each case is down to about 80 occurrences per 100 recorded instances, while the minor spelling alternative has advanced to about 20 instances per 100 cases. Why is this so?

Before trying to answer that question, however, let us perform two additional frequency studies, exploring further restriction possibilities offered by Google. The purpose of this measure is simply to try to determine how alternative ways of slicing the Web might possibly influence our results in this context. Hence, Table 5 presents data from searches on the same spelling variants but now constrained through Google's domain restriction facility (January 2004):

Table 5: Comparison of the frequencies of the two spelling variants *colour* and *color* in language material published within the (primarily) American domain .edu and the British domain .ac.uk., as mined by Google

	.edu	%	.ac.uk	%
colour	136 000	4.1	318 000	83.5
color	3 150 000	95.9	62 800	16.5
Total	3 286 000	100	380 800	100

Changing the search denominator to the general educational domains of the US and the UK, respectively, takes us down to the four million range of matches. Evidently, this manner of delimitation brings the American-based figures somewhat closer to those obtained earlier from the sampled corpora, while the UK-based figures remain more or less the same as those derived previously through Google's language variety restriction facility. Clearly, this divergent outcome warrants a further curb on the Web material in order to see if the new pattern is persistent, or if it is rather to be attributed to possible heterogeneity within the educational domains. In Tables 6 and 7 this aim is realized as an even more constrained search of the target spelling variants, now within the range of four individual newspapers only, two American and two British:

Table 6: Comparison of the frequencies of the two spelling variants *colour* and *color* in the American papers *The New York Times* (nytimes.com) and *The Washington Post* (washingtonpost.com), as mined by Google

	New York Times	%	*Washington Post*	%
colour	23	1.4	41	1.4
color	1 580	98.6	2 950	98.6
Total	1 603	100	2 991	100

Table 7: Comparison of the frequencies of the two spelling variants *colour* and *color* in the British papers *The Times* (thetimes.co.uk) and *The Guardian* (guardian.co.uk), as mined by Google

	The Times	%	*The Guardian*	%
colour	686	98.8	10 500	97.3
color	8	1.2	293	2.7
Total	694	100	10 793	100

Bringing us down to the 16 000-word range of matches, this last restriction in terms of newspapers yields relative figures which are compatible with those reported from standard corpora. In other words, this is to say that the investigated papers collectively returned the chiasmic usage pattern observed previously, with a *colour/color* ratio of almost 1/99 for the American ones and a ratio of approximately 98/2 for the British ones. We may also note, incidentally, that the majority of matches here (10 793) derive from one single paper, *The Guardian*, whose textual contribution to the Web thus seems to be almost on a par with the size of the BNC (11 460 matches).

Why is it, then, that only domain-restricted searches yield results which are compatible with those from standard corpora? Arguably, there are at least two related reasons for this outcome. One is that standard corpora often contain a great deal of newspaper text, which is easy to access for people who are in the process of compiling corpora, thus setting journalistic language as a kind of stylistic denominator for this type of sampled material. To a certain degree, this might explain the greater output parallelism achieved between corpora and Web-based material when searches are restricted to media-based domains in general and newspapers in particular. Another, more significant reason is the uncontrolled, heterogeneous character of the Web. Featuring not only a large body of conventional manifestations of language, primarily of the commercial type (cf. Lawrence and Giles 1999), but also a wide range of "uninformed, colloquial, provisional and improvised language use of the spontaneous kinds encouraged particularly in chat rooms and news lists" (Renouf 2003), the Web may be said to represent, on the whole, a rather different type of textual data than standard corpora. Specifically, the large unfiltered component of texts and text fragments (cf. Meyer et al. 2003), including various types of errors and awkward constructions, gives it an informal bent which renders comparisons

with sample-based collections more difficult, largely because the latter type of material tends to be more balanced, filtered and homogenous. The corollary of this observation is that, looking into the Web *in toto*, we should not even expect the results to be the same as those derived from standard corpora, simply because these collections of material are composed differently and represent different stylistic levels. It is only when we can restrict our searches of the Web to specific, controlled domains that we can achieve frequency-based comparability. This also corroborates claims by previous scholars, e.g. those of the WebCorp team, that while it is generally difficult to use the Web in its entirety for frequency studies or delicate text analysis, slices of it in terms of domain-specific searches are more rewarding as they lead to a higher level of precision (e.g. Renouf 2003, Renouf et al. 2004).

5 Exploring Google and WebCorp further

To bring the discussion to a close now, let us return to the word *Taliban*. We used this word previously as a good example of when it is particularly profitable to use the Web to gather linguistic information, simply because standard corpora are not recent enough to include it, at least not in sufficient number. In that context, it was also implied that the grammatical properties of this word are of specific interest, something which is based on an apparent concord difference between American and British English (cf. Ohlander and Bergh 2004).

Using the domain restriction facility of Google, specifically the respective commercial domains .co.uk and .com, let us see what the engine has to say on this point. The search strings in question were randomly designed to include the phrase *the Taliban* followed by the singular or plural forms of three high-frequency verbs in English, *say*, *have* and *be* (Table 8):

Table 8: Absolute distribution of different concord patterns with the word *Taliban* in the .co.uk and .com domains, as mined by Google

Search string	.co.uk	.com
the Taliban say	54	815
the Taliban says	35	747
the Taliban have	1 010	9 300
the Taliban has	556	8 270
the Taliban were	917	7 730

the Taliban was	431	7 180
total: *the Taliban* + singular verb	1 022	16 197
total: *the Taliban* + plural verb	1 981	17 845

As we can see, there is a fairly clear trend for plural concord to be more common with the word *Taliban* in British English (1981 vs. 1022 matches), whereas the usage seems to be more equally divided in American English (16 197 vs. 17 845 matches). However, since these commercial domains are very comprehensive and rather mixed with regard to language, it might be worth while to restrict the search further, using newspaper language proper as the target domain. The results of this search are presented in Table 9, mined by Google but restricted via WebCorp to 120 Web pages.

Table 9: Absolute distribution of different concord patterns with the word *Taliban* in British and American newspapers, as mined by Google/WebCorp

Search string	UK newspapers	US newspapers
the Taliban say	17	3
the Taliban says	5	6
the Taliban have	155	52
the Taliban has	120	102
the Taliban were	137	59
the Taliban was	101	77
total: *the Taliban* + singular verb	226	185
total: *the Taliban* + plural verb	309	114

Here we can see that the pattern with plural concord persists for British English (309 vs. 226 matches), while that for American English is now refined to show that it is in fact singular concord which is typical for this transatlantic variety (185 vs. 114 matches). Consequently, these results confirm the findings of a recent study of the grammatical behaviour of *Taliban*, based on a variety of different textual sources, where it was shown that British English typically favours

the plural form and American English the singular form (Ohlander and Bergh 2004). Again, then, we seem to have confirmation of the pattern that domain-specific searches are more reliable than overall searches of the Web, and that the more well-defined the domain, the more clear-cut the frequency results.

Adding a bit of culinary zest to the present paper, finally, we may sum up this outing into Web linguistics, with its particular focus on large-scale language collections and state-of-the-art search technology, by a short concluding statement which seems to capture the essence of the discussion: the Web is best enjoyed in carefully cut slices, preferably based on the raw capacity of Google and spiced according to taste with the fine-tuned linguistic facilities of WebCorp.

References

Bergh, Gunnar, Aimo Seppänen and Joe Trotta. 1998. Language corpora and the Internet: A joint linguistic resource. In A. Renouf (ed.). *Explorations in Corpus Linguistics*. Amsterdam: Rodopi, 41–54.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Botley, Simon, Anthony McEnery and Andrew Wilson (eds.). 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi.

Brin, Sergey and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. Available at http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.

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

Granger Sylviane. 2004. Computer learner corpus research: Current status and future prospects. In U. Connor and T. Upton (eds.). *Corpus Linguistics in North America 2002*. Amsterdam: Rodopi, 123–145.

Grefenstette, Gregory and Julien Nioche. 2000. Estimation of English and non-English language use on the WWW. *Proceedings of the Recherche d'Informations Assistée par Ordinateur* (RIAO). Paris, 237–246.

Ide, Nancy, Randi Reppen and Keith Suderman. 2002. The American National Corpus: More than the Web can provide. *Proceedings of the Third Language Resources and Evaluation Conference* (LREC), Las Palmas, Spain, 839–844.

Kilgariff, Adam. 2001. Web as corpus. In P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.). *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster: UCREL, 342–344.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the special issue on the Web as corpus. *Computational Linguistics* 29: 333–347.

Lawrence, Steve and Lee Giles. 1999. Accessibility of information on the Web. *Nature* 400:107–109.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics*. Second edition. Edinburgh: Edinburgh University Press.

Meyer, Charles, Roger Grabowski, Hung-Yul Han, Konstantin Mantzouranis and Stephanie Moses. 2003. The World Wide Web as linguistic corpus. In P. Leistyna and C. Meyer (eds.). *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 241–254.

Ohlander, Sölve and Gunnar Bergh. 2004. *Taliban* – a rogue word in present-day English grammar. *English Studies* 85: 206–229.

Renouf, Antoinette. 2003. WebCorp: Providing a renewable data source for corpus linguists. *Language and Computers* 48: 39–58.

Renouf, Antoinette, Andrew Kehoe and David Mezquiriz. 2004. The accidental corpus: Some issues in extracting linguistic information from the Web. *Language and Computers* 49: 403–419.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Xu, Jack. 2000. Multilingual search on the World Wide Web. *Proceedings of the Hawaii International Conference on System Science*. Maui, Hawaii.