The World Wide Web as Linguistic Corpus

Charles F. Meyer, Roger Grabowski, Hung-Yul Han, Konstantin Mantzouranis and *Stephanie Moses*

The University of Massachusetts Boston, USA

Abstract:

Increasingly, corpus linguists have begun using the World Wide Web as a corpus for conducting linguistic analyses. The Web, however, is really a very different kind of corpus: we do not know, for instance, precisely how large it is or what kinds of texts are on it. In this chapter, we evaluate the Web as a linguistic corpus, providing estimates of its size and composition. In addition, we conduct a series of sample analyses of the Web, demonstrating that while commonly available search engines have definite limitations, they can in a matter of seconds retrieve extremely large volumes of data that are very relevant to a corpus analysis, and also provide frequency information that may not be entirely accurate but suggestive of how frequently particular words and grammatical constructions occur.

1. Introduction

As the World Wide Web has grown in size and popularity, linguists and language teachers have turned to it as a valuable resource for both studying and teaching the structure of English and other languages. The WebCorp Project at the University of Liverpool has developed a Web-based concordancer that can search for lexical items on the Web and display them in KWIC format.[1] KWiCFinder is a pc-based program that submits various kinds of searches to AltaVista and then provides an interface for reviewing the results in KWIC format.[2] The Division of English as an International Language at the University of Illinois-Champaign/Urbana has developed a teaching strategy for non-native learners of English -- "TheGrammar Safari --that has students engage in data-driven learning activities that can be done on the Web using commonly available search engines.[3]

The easy access to the Web makes it an attractive resource for conducting corpus analyses, but there is a key difference between the Web and other more commonly used corpora: while we know precisely what we're analysing in a corpus such as the British National Corpus, when conducting a search on the Web, we have no idea what kinds of texts our search results have been taken from. This difference raises an obvious methodological question: can we 'trust' the results that we obtain from any linguistic analysis of the Web? To begin answering this question, we discuss whether it is possible to gauge the size of the

[1] See http://www.webcorp.org.uk/

[2] See http://miniappolis.com/KWiCFinder/KWiCFinderHome.html

[3] See http://deil.lang.uiuc.edu/web.pages/grammarsafari.html

Web and the rate at which its size is increasing, to determine with any certainty the kinds of texts that exist on the Web, to calculate which portion of the Web commonly available search engines actually obtain results from, and to describe the very real limitations of search engines in terms of what they can or cannot search for. In addition, we will report the results of some sample linguistic analyses we conducted and describe the implications of our searches for linguistic description and teaching.

2. The size and composition of the Web

The Web is the ultimate ‘monitor' corpus (Sinclair 1991): texts go in and out of it, making it a very dynamic linguistic corpus. However, the Web is different from other monitor corpora, such as the Bank of English Corpus, because we do not know its precise size or the kinds of texts that comprise it.

There have been various attempts to estimate the size of the Web by calculating the number of ‘discrete' Web pages that exist. Lawrence and Giles (1999) conducted a study between February and March 1999 and concluded that there were 800,000 Web pages. A study in 2000, conducted by Inktomi and the NEC Research Institute, claimed that there were over one billion “unique” Web pages. As of November 2002, the search engine Google had indexed over three billion Web pages.[4]

These varying estimates reflect both the difficulty of estimating the size of the Web and the fact that in the last couple of years the size of the Web has increased significantly. To illustrate just how quickly the Web has grown, in Table 1, we provide the number of hits for the words *chairman*, *chairperson*, and *chairwoman* that the search engine Google returned between 2000 and 2002.

Table 1: Number of word hits between 2000-2002

	chairperson	*chairman*	*chairwoman*
September 2000	565,000	4,680,000	88,000
February 2001	766,000	5,810,000	103,000
February 2002	1,110,000	7,540,000	130,000

The figures in Table 1 illustrate a steady growth in the Web, with the number of hits for each lexical item nearly doubling between September 2000 and February 2002. The large number of hits for each item also reveals the sheer magnitude of the size of the Web, a size that is obscured by counting the number of pages rather than the number of words of text on the Web.

To obtain an estimate of the number of words on the Web, it is instructive to extrapolate from calculations made in Lawrence and Giles (1999), even though

[4] See http://www.google.com/

these calculations are currently fairly out of date. Lawrence and Giles (1999: 107) examined a number of Web pages and calculated that on average a Web page contained 7.3 kb of text (with html formatting and white space excluded from this calculation). In Table 2, we calculate just how large a corpus of 800,000 Web pages would be.

Table 2: The size of the Web

Kilobytes (Kb)	6,442,450,944
Megabytes (Mb)	6,291,456
Gigabytes (Gb)	6,144
Terabytes (Tb)	6
Number of Words	836,070,780,477

The total word length in Table 2 indicates that the Web is huge -- far larger than any currently existent corpus. And since the Web has grown considerably since Lawrence and Giles' (1999) study, it can be assumed that the Web is probably well over a trillion words in length. There is thus ample text on the Web for linguistic analysis. The challenge, though, is determining what kinds of texts are on the Web -- the particular genres that they represent -- and once this determination is made, figuring out how to gain access to whatever linguistic information one is seeking.

Even though the Web is multi-lingual, it consists predominantly of texts in English. Working from various databases on Web page content, Pleasants (2001) came up with the following breakdown of the percentage of Web pages in various languages:

English (68.4%)
Japanese (5.9%)
German (5.8%)
Chinese (3.4%)
French (3.0%)
Spanish (2.4%)
Russian (1.9%)
Italian (1.6%)
Portuguese (1.4%)
Korean (1.3%)
Other (4.9%)

Ironically, even though most Web pages are in English, the majority of individuals using the Web (52.5%), according to Pleasants (2001), are native speakers of languages other than English. It will undoubtedly be the case that as more of the non-English speaking world gains access to the Internet, a higher percentage of languages other than English will be represented on the Web. Nevertheless, the Web is primarily a corpus of English.

To determine what kinds of English are represented on the Web has proven a difficult task. Lawrence and Giles (1999) give a general breakdown of the percentages of pages investigated in their study devoted to a particular content area:

Commercial	(83%)
Scientific/Educational	(6%)
Health	(3.8%)
Personal Web pages	(2.2%)
Societies	(2%)
Pornography	(1.5%)
Community	(1.4%)
Government	(1.1%)
Religion	(.9%)

Even though the majority of Web pages contain some kind of commercial content, many pages containing the .com domain name would include texts commonly found in corpora, such as newspaper articles. The other content areas listed above would also contain texts of interest to corpus linguists: scholarly articles of a scientific or educational nature, government documents, popular discussions of health topics, even texts taken from personal home pages, since as we will show in later sections, the language used on these pages provides an ‘unfiltered' and ‘unedited' view of actual language usage.

3. Conducting linguistic analyses of the Web

Because the Web is a huge, grammatically unanalysed corpus, obtaining information from it raises the same problems that analysing any lexical corpus will raise: the appropriate search tools for the linguistic constructions being studied must be selected; once results are obtained, they must be subjected to some kind of quantitative analysis; if a given construction does not have a unique linguistic form (e.g., if a lexical item is polysemous), those constructions under investigation must be distinguished from those forms not relevant to the analysis; if a grammatical construction, rather than a lexical item, is being studied, a lexical item associated with the grammatical construction must be included with any search for the construction because manual analysis of the Web is time-consuming and often unproductive. Because of the size of the Web, these problems are compounded. In this section, we carry out a series of linguistic analyses to demonstrate how various linguistic constructions -- both lexical and grammatical -- can be studied on the Web.

3.1 Selecting the appropriate search tool

The most challenging aspect of any corpus analysis is quickly and efficiently locating the relevant linguistic constructions being studied. As we noted earlier, there are both Web-based and pc-based programs designed specifically to search the Web and produce KWIC concordances. The WebCorp concordancer, for instance, allows one to submit searches to a search engine, such as Google or AltaVista, and then displays the results of the search in a KWIC format, providing not just examples of the particular item being searched for, but also hyperlinks to the Web pages containing the examples. Although Web-based programs such as this provide very good interfaces for linguistic analysis, they are slower than the major search engines and their coverage of the Web is limited. In a search on 29 March 2002 for instances of *chairperson*, the WebCorp concordancer took 2.5 minutes (with a high speed Internet connection) to return 75 examples of this word. A similar search on Google took .20 seconds to locate 1,190,000 Web pages containing *chairperson.*

Because Google is a general purpose search engine, it lacks the useful interface for linguistic analysis that the WebCorp concordancer has. But Google does provide short excerpts in which search items appear as well as hyperlinks to the Web pages containing the items. In addition, Google' s speed and coverage are very important for linguistic analysis of the Web because searches often need to be restructured multiple times to achieve the desired results, and this kind of searching is best achieved with a search engine that is fast. Web-based linguistic resources such as the WebCorp concordancer will undoubtedly improve over time, but from our experience, we have found that search engines such as Google are most useful for conducting linguistic analyses of the Web.

Another advantage of using a search engine such as Google is that it covers much more of the Web than the other major search engines. A common complaint about search engines that is repeatedly echoed in the literature is that their coverage is inadequate and that they miss much of the information that exists on the Web, such as information available in the 'deep' Web: that part of the Web, as Bergman (2001) indicates, which escapes detection by search engines. However, even though search engines may cover only a fraction of the Web pages that actually exist, they nevertheless cover enough of the Web to yield more than enough linguistic information. In fact, the real challenge in conducting searches is sorting through an excessive amount of linguistic detail that search engines typically return to isolate the relevant linguistic information being sought.

Sullivan (2001) provides useful statistics on the amount of the Web that the more popular search engines are able to reach (see Table 3).

Table 3: Web page coverage by various search engines (as of 11 December 2001)

Search Engine	**# of Pages Indexed**
Google	1.5 billion (but able to reach an additional 500 million pages not indexed)
Fast	625 million
AltaVista	550 million
Northern Light	390 million

Table 3 indicates that Google reaches significantly more Web pages than any of the other search engines, though as we will demonstrate in the next section, most of the major search engines reach enough of the Web to yield fairly comparable information.

3.2 Evaluating frequency information generated by search engines

In addition to returning items being searched for, search engines will provide statistics on the number of ‘hits' that a particular search item yields. Because frequency information is so important to corpus linguists, it is worth looking in some detail at the frequency information that search engines provide. Table 4 lists the number of ‘hits' returned on three different occasions by the search engines in Table 3 for three lexical items: *chairman*, *chairperson*, and *chairwoman.*

Table 4: Consistency of frequency counts generated by search engines

	Google	**Fast**	**AltaVista**	**N. Light**
29-3-2002				
chairman	7,710,000 (85%)	4,325,141 (86%)	3,506,210 (87%)	2,063,336 (95%)
chairperson	1,190,000 (13%)	644,491 (13%)	444,227 (11%)	30,017 (1%)
chairwoman	141,000 (2%)	63,735 (1%)	65,312 (2%)	80,495 (4%)
Total	9,041,000	5,033,367	4,015,749	2,173,848
4-4-2002				
chairman	7,710,000 (85%)	4,325,166 (86%)	3,020,596 (87%)	2,081,540 (95%)
chairperson	1,190,000 (13%)	644,491 (13%)	391,205 (11%)	30,255 (1%)
chairwoman	141,000 (2%)	63,876 (1%)	59,667 (2%)	80,733 (4%)
Total	9,041,000	5,033,533	3,471,468	2,192,528
15-5-2002				
chairman	7,810,000 (86%)	6,585,252 (86%)	3,023,125 (87%)	2,156,782 (95%)
chairperson	1,170,000 (13%)	955,564 (12%)	391,262 (11%)	31,176 (1%)
chairwoman	147,000 (2%)	116,245 (2%)	59,786 (2%)	82,789 (4%)
Total	9,127,000	7,657,061	3,474,173	2,270,747

Before commenting on the information in Table 4, it is important to define exactly what a ‘hit' is. Search engines (with the exception of AltaVista) do not count how frequently a given search term occurs on the Web but instead provide counts of the number of pages actually found containing the term. Thus, if a given term occurs three times on a particular Web page, the search engine will report a frequency of one, not three. In addition, overall frequencies can be deceptive because a given Web page may be mirrored on more than one Web site. Consequently, if the same Web page exists on multiple servers, a search engine will count each Web page as a separate hit, even though the Web pages contain identical information.

With these caveats in mind, Table 4 reveals that three of the four search engines returned fairly similar frequency information: the percentages for each lexical item were similar over time, even though the raw frequencies were quite different. The notable exception is the search engine Northern Light, whose frequencies were very different from the other search engines.

Although all search engines will return page hits, only AltaVista will provide string frequencies as well. Table 5 compares the number of hits with the number of string frequencies for the three lexical items.

Table 5: 'Hits' versus 'string frequencies' from AltaVista (15 May 2002)

Lexical Items	**Hits**	**Strings**
Chairman	3,023,125 (87%)	6,015,339 (86%)
Chairperson	391,262 (11%)	904,303 (13%)
Chairwoman	59,786 (2%)	81,026 (1%)
Total	3,474,173	7,000,668

Not surprisingly, Table 5 demonstrates that for any given search item, there will be many more strings than hits. However, when percentages are considered rather than raw figures, the results are fairly similar. We can therefore conclude that search engines provide at least a rough guide to the relative frequency of a given linguistic construction. It is probably safe to assume that even though the gender-neutral term *chairperson* is being used, it is new enough to not have replaced the more well established term *chairman*. And the low frequency of *chairwoman* might simply reflect the fact that fewer females are in positions where this term would be appropriate. But the true test of whether *chairman* holds sway would be to compare its frequency of usage with that of the term *chair*, a term whose multiple meanings complicates this type of comparison.

3.3 Searching for polysemous lexical items

Polysemous lexical items pose problems for any corpus analysis. In a smaller corpus, one could simply go through each lexical item by hand, removing those items whose meaning is not relevant to the analysis. But the size of the Web makes such a strategy impractical. However, if one includes in the search additional search terms, it is possible to obtain information on polysemous lexical items.

Table 6 lists the results of a search not just for *chairman*, *chairperson*, and *chairwoman*, but for *chair* as well. To insure that the 'furniture' meaning of *chair* was excluded from the search, the additional search term *committee* was included with each of the words in Table 6 to narrow the range of items returned with the meaning of 'furniture' and to permit valid comparisons between each of the expressions.

Table 6: Frequency of *chairman*, *chairperson*, *chairwoman*, and *chair* occurring with the word *committee* (Google, 9 August 2002)

Search Item	**Frequency**
chairman	2,440,000 (43%)
Chair	2,660,000 (47%)
chairperson	526,000 (9%)
chairwoman	66,200 (1%)
Total	5,692,200

Interestingly, the results in Table 6 indicate that the gender neutral term *chair* is gaining popularity over the other choices. However, without going through each hit for *chair* individually, it is difficult to know how many instances of *chair* are nouns rather than verbs, as in example (1), a sentence that ironically contains *chairman* to designate the individual doing the chairing:[5]

(1) Steven B. Solomon, *Chairman* and CEO of CT Holdings, Inc., (OTCBB: CITN), which develops and markets the Citadel Technology™ line of network security and privacy software, announced today that he will *chair* the Committee on Computer Privacy and Data Security Standards. (emphasis added) (http://www.ct-holdings.com/Press%20Releases/press010924.htm)

Nevertheless, the counts do suggest that those seeking a gender neutral equivalent to *chairman* are more likely to choose *chair* than *chairperson.* And in cases where one does wish to indicate the gender of the individual, *chairman*, a very well established word, is much more likely to be chosen than *chairwoman.*

We encountered a similar situation in comparing the use of *hopefully* with two synonymous equivalents: *I hope* and *It is hoped.* Table 7 gives results of a search for these three expressions.

[5] In this and subsequent examples, we have italicised the parts of example sentences under discussion.

Table 7: Comparison of *hopefully* with *I hope* and *It is hoped* (Google, 10 July 2002)

Search Item	**Frequency**
hopefully	3,450,000 (45%)
I hope	3,410,000 (45%)
It is hoped	684,000 (10%)
Total	7,544,000

We examined manually the first several hundred hits for *hopefully* to determine how many of the hits were for the manner adverbial sense of *hopefully* (meaning ‘in a hopeful manner', as in example 2) and how many were for *hopefully* as an attitudinal disjunct (meaning ‘I hope', as in example 3).

(2) Then one morning the sky was overcast. We waited *hopefully*. Then the rain fell.
http://www.amishnews.com/amisharticles/amishspeak.htm

(3) *Hopefully*, future Web innovations will emulate the example set by the Web Consortium in its work on CSS.
http://www.useit.com/alertbox/styles

Surprisingly, only a few instances of *hopefully* were used as a manner adverbial, suggesting that the vast majority of the frequency counts for *hopefully* in Table 7 reflect the use of this expression as an attitudinal disjunct. Thus, we can conclude from Table 7 that even though *hopefully* as a disjunct was at one time a highly stigmatised usage, it is now used as commonly as its periphrastic counterpart, *I hope*. And the highly impersonal *It is hoped* is much rarer, largely because the informal contexts in which the three expressions are likely to be used make the passivized construction *It is hoped* an inappropriate choice.

In other cases, however, search results are better suited to qualitative rather than quantitative analysis. Table 8 contains the results of a search for the lexical items *disinterested* and *uninterested*. We wanted to investigate the extent to which *disinterested* was used to mean ‘uninterested' rather than its more traditional meaning of ‘impartial'. Although the frequency information in Table 8 suggests that *disinterested* is more common than *uninterested*, as we began examining the hits that were returned, we encountered many examples casting doubt on the reliability of the frequency results for *disinterested*.

Table 8: The frequency of *disinterested* and *uninterested* (Google, 10 July 2002)

Search Item	**Frequency**
disinterested	163,000 (67%)
uninterested	81,800 (33%)
Total	244,800

First, many of the hits were from sites giving prescriptive information on the usage of *disinterested* to mean 'uninterested'. Thus, these sites do not provide reliable information on the actual usage of *disinterested.* More importantly, we noticed a fairly even mixture of examples where *disinterested* meant either 'uninterested' or 'impartial'. To obtain any reliable quantitative information would therefore require going through each hit individually, an analysis that would require too much time to make it worthwhile.

In examining examples, however, we noticed that many of the instances where *disinterested* meant 'impartial' (see examples 4-6 below) came from sites offering some kind of legal advice, suggesting that this usage may be more confined to legal usage.

(4) If the trust instrument required a *disinterested* trustee or a *disinterested* trustee is required to carry out provisions of the trust, an interested person or the resigning trustee should petition the court for the appointment of a successor.
http://courts.co.calhoun.mi.us/epic0252.htm

(5) Do not proceed until your *disinterested* third party or state-approved monitor is present.
http://www.abcselfstudy.com/Forms/dtp.htm

(6) In most states, a formal will must be written, signed by the person making the will and signed by two or more *disinterested* witnesses.
http://www.gottrouble.com/legal/estate_planning/legally_valid_wills.html

On the other hand, when *disinterested* was used to mean 'uninterested', the context of usage was clearly less formal, as examples (7) and (8) demonstrate:

(7) Ideally, no one would ever unsubscribe from our email publications, and everyone would stay interested. But this is never the case, and we as publishers should make it easy for our subscribers to remove themselves should they become *disinterested* in the content.
http://ezine-tips.com/articles/strategy/20000912.shtml

(8) Are young people *disinterested* in using the internet for more socially aware needs?
http://youthlink.takingitglobal.org/express/article.html?cid=21&pn=4

While in cases such as this the Web cannot be used to uncover quantitative information on usage, it can be used to provide examples that are suggestive of the contexts in which the usage predominates.

3.4 Searching for syntactic constructions

The sheer size of the Web ironically makes it easier to isolate certain usages of syntactic constructions, since in addition to searching for lexical items associated with a particular syntactic construction (such as the relative pronoun *who*), one can include other items that more precisely specify a particular syntactic construction. For instance, we investigated the extent to which the relative pronoun *who* was used where normally one would expect *whom.* A simple search for *who* or *whom* will yield millions of hits, and without inspecting each hit individually, it will be impossible to determine whether *who* or *whom* is functioning as a subject or an object. However, if a specific transitive verb is included in the search along with a commonly occurring subject (e.g. the first person pronoun *I* in a constructions such as *who I like),* it will be possible to determine how frequently *who* rather than *whom* is used when the relative pronoun is functioning as object.

Table 9 contains the results of a search for relative clauses headed by *who* or *whom* that contain the pronoun *I* functioning as subject and a series of transitive verbs: *like, know, called, gave,* and *took.*

Table 9: The frequency of *who* and *whom* with select transitive verbs (Google, 10 July 2002)

Search Item	**Frequency**
who I like	7,700 (67%)
whom I like	3,800 (33%)
Total	11,570
who I got to know	422 (45%)
whom I got to know	522 (55%)
Total	944
who I called	1,370 (42%)
whom I called	1,870 (58%)
Total	3,240
who I gave	1,100 (28%)
whom I gave	2,810 (72%)
Total	3,910
who I took	1,160 (27%)
whom I took	3,150 (73%)
Total	4,310

The results in Table 9 are anything but straightforward and support Huddleston and Pullum's (2002: 464) assertion that while 'In short simple constructions…there is a rather sharp contrast between *who* and *whom*.…It would be a mistake to say, however, that *whom* is confined, or even largely confined, to formal style'. They note, for instance, that even a fairly colloquial construction such as *who(m) I got to know* will not necessarily always elicit *who*, as illustrated by the results in Table 9 and the informal flavour of examples (9) and (10):

(9) Realistically, it is unlikely that all 400-plus of us will be in each others' address lists, but even those *whom I got to know* less well, or even just exchanged greetings with me in the hallways will remain a part of my UMBS experience, and hopefully get in touch if they are ever in my neck of the woods in years to come.
http://themsj.com/news/241865.html

(10) This site has been built by a good friend of mine, *who i got to know* through Tarantulas.com and the chat room on this site.
http://www.geocities.com/coollinks2/pets/pets.html

There is only a consistent preference for *whom* over *who* when the relative pronoun heads a clause containing a verb, such as *gave* or *took*, that regularly occurs with a preposition. In (11), *took* elicits the preposition *from:*

(11) If I have any duty in the matter, it is to give back the money to those *from whom* I took it; not to pay it over to such villains as you.
http://www.fourmilab.ch/etexts/www/NoTreason/NoTreason_chap12.html

Whom is obligatory in (11) because it immediately follows *from;* if *from* were stranded (e.g., *…to those whom I took it from….*), *who* would also be possible. But because so many instances of *gave* contained the preposition *from* positioned before the relative pronoun (750 out of 3,150 instances, or 24%), *gave* will naturally occur more frequently in relative clauses with the relative pronoun *whom.*

Further variation in the use of *who* or *whom* can be found in examples containing comment clauses such as *I believe* -- clauses that often lead to ‘hypercorrection', and the use of *whom* even when the relative pronoun is functioning as subject of its clause (Quirk et al. 1985: 368, note [a]; Huddleston and Pullum 2002: 466). Table 10 gives figures for the frequency of the constructions *who I believe is* and *whom I believe is.*

Table 10: The frequency of *who* and *whom* in comment clauses (Google, 10 July 2002)

Search Item	**Frequency**
who I believe is	4,820 (82%)
whom I believe is	1,030 (18%)
Total	5,850

Although there is a clear preference for the use of *who* in constructions of this type (12), *whom* nonetheless does occur (13), providing further evidence that the status of *whom* in Modern English remains quite variable.

(12) If I know someone *who I believe is* no longer able to drive safely, what do I need to do to get the driver retested or checked on?
http://www.dor.state.mo.us/mvdl/drivers/unsafe.htm

(13) I am obviously too young to know anyone in this photo, other than the young mascot *whom I believe is* Tony Surman.
http://homepage.powerup.com.au/~woomera/sp_01.htm

Biber et al. (1999: 610-11) provide corpus findings demonstrating that in a range of different genres -- conversation, fiction, news, and academic prose -- the relative pronoun *whom* is much rarer than *who*. Our study confirms these results, but at the same time it illustrates that in certain contexts (e.g., following a preposition and with a range of transitive verbs), *whom* will be preferred to *who*. Thus, we would agree with Huddleston and Pullum' s (2002) claim that the choice between *who* and *whom* involves more than simply a choice between formal and informal style.

4. Conclusions

We have shown in this paper that the Web can yield valuable information, even though its size and the particular kinds of texts on it are difficult to estimate. Although frequency information generated by search engines must be interpreted with caution, such information is ‘suggestive' and can give a sense of which linguistic usages are common and which are not. In addition, the examples that can be found on the Web are valuable for establishing common patterns of usage. We were particularly struck in our analyses by the ‘unfiltered' nature of the Web: much of the data we encountered in our analyses was unedited and thus reflective of how people actually use language.

The challenge for corpus linguists in the future will be to develop tools that will not only help linguists find linguistic constructions on the Web, but enable them to locate these constructions within particular genres. At present, it is only possible to do random searches. We can only hope that in the future those creating Web pages will make greater use of ‘meta tags': tags that are inserted into a document annotated with html or xml markup that provide descriptive information about the content of a document, information that many search engines can search for. There are also linguistic initiatives, such as the Open Language Archives Community, proposing standards for the annotation of electronic texts -- annotation called ‘meta-data', which would include such information as the ‘source' and ‘subject' of a document. And as more linguistic corpora become available on the Web, such annotation will allow one to be more specific when searching for linguistic constructions in a particular genre.

References

Bergman, M.K. (2001). ‘The deep Web: Surfacing hidden value'. *The Journal of Electronic Publishing,* 7:
http://www.press.umich.edu/jep/07-01/bergman.html

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999). *Longman grammar of spoken and written English.* Harlow, Essex: Longman.

Huddleston, R. and G.K. Pullum (2002). *The Cambridge grammar of the English language.* Cambridge: Cambridge University Press.

Lawrence, S. and C.L. Giles (1999). ‘Accessibility of information on the Web'. *Nature*: 107-9.

OLAC: Open Languages Archive Community. http://www.language-archives.org/

Pleasants, N. (2001). ‘Languages of the Web'. *ClickZ Today*, May 11, 2001. http://www.clickz.com/int_mkt/global_mkt/article.php/841721

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985). *A grammar of contemporary English.* London: Longman.

Sinclair, J. H. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sullivan, D. (2001). Search engine sizes. http://www.searchenginewatch.com/reports/sizes.html

Inkotomi Corporation (2000). ‘Web surpasses One billion documents'. Press Release: http://www.inktomi.com/new/press/2000/billion.html