The Web as corpus versus traditional corpora: their relative utility for linguists and language learners

Robert Lew

> *The Web, teeming as it is with language data, of all manner of varieties and languages, in vast quantity and freely available, is a fabulous linguists' playground.*
>
> - Adam Kilgarriff and Gregory Grefenstette (2003: 333)

1. Introduction[1]

Electronic corpora of natural language have grown dynamically in the recent decades: in their number, volume, as well as in importance (Biber et al. 1998; Sinclair 1991; Walter and Harley 2002). The population of typical corpora users is no longer restricted to the inner circles of lexicographers, linguists and experts in Natural Language Processing (including speech and character recognition, machine translation, spellchecking and grammar checking). Increasingly, corpora are being embraced by representatives of the less esoteric and less technical language-related professions, such as translators and language teachers, but also by language learners themselves (Aston 1997b; Tribble 1991; Varantola 2003). It is a sign of the times that corpus samples have even made their way into learners' dictionaries (e.g. Sinclair 1995).

On the other hand, the World Wide Web, the hypertext, multimedia section of the internet,[2] generally assumed to have originated in 1994 (de Schryver 2002), is developing with such momentum that in some areas of applications it may encroach on the niches freshly filled by 'traditional' electronic text corpora. A number of authors have proposed to treat and use the textual content of the world's web pages as a corpus (Fletcher 2004; Grefenstette 1999; Kilgarriff and Grefenstette 2003; Resnik and Smith 2003; Rundell 2000; de Schryver 2002;

[1] I am indebted to Przemysław Kaszubski and Włodzimierz Sobkowiak for their helpful comments on an earlier version of this chapter.

[2] In its everyday sense, the internet is understood in the technically narrower sense of the World Wide Web.

Smarr and Grow 2002). As Kilgarriff and Grefenstette rightly point out, '[l]anguage scientists and technologists are increasingly turning to the Web as a source of language data, because it is so big, because it is the only available source for the type of language in which they are interested, or simply because it is free and instantly available' (Kilgarriff and Grefenstette 2003: 333).

In the present chapter I will try to discuss the usefulness of the two types of resources (i.e. traditional text corpora and the WWW) for two broad categories of users (and uses). Firstly, I want to evaluate their usefulness in the context of foreign language teaching, seen as a branch of applied linguistics, as a potential tool for solving ad hoc lexical queries. I will not address the (no doubt very interesting and important) issues of the role played by corpora in the creation of learning tools and aids, such as textbooks, grammar books, or dictionaries (e.g. Aston 1997a; Aston 1997b; Partington 1998; Willis 2000); or the direct use of corpora for inductive data-driven learning (e.g. Johns 1991). The scope of my discussion will also be restricted to the English language and to the lexical dimension. The other category of user that I want to focus on here will be the professional linguist. Here, I would like to focus primarily on the use of such textual resources for the verification of linguistic hypotheses by those linguists of various specialisms who see corpora as important sources of linguistic data, that is those who accept the empirical methodology to a lesser or greater degree, and who see texts produced in a language as a valid source of data for linguistic enquiry. This type of fairly general linguistic application appear to be of most general interest to a broad range of linguists, as opposed to certain more specialized applications relevant to, for example, computational linguists. In what follows, I will try to trace where the expectations and needs of the linguist-researcher and language learner converge, and where they diverge.

Given the two types of corpus-like resources (traditional corpora and the WWW), and the two broad categories of their users outlined above, it seems that a further distinction would usefully be made, based on the mechanism of accessing the resources in question. This is so

because the access mechanism used has a significant impact on the functional qualities of the resources, and thus their practical utility. I believe it is necessary to consider at least three different access mechanisms that the user (the practising linguist or the language learner, respectively) can employ to communicate with the textual database. The three mechanisms are: a dedicated concordance application running on the user's desktop computer; a server-based concordance application accessed through the hypertext protocol; and a publicly accessible general-purpose search engine.

In the present chapter I will try to compare the usefulness of traditional corpora and the-World Wide Web-as-corpus with reference to the following criteria: size of the resources; linguistic representativeness; balancing and noisiness; functionality and access mechanism.

2. Size of textual resources

The resource that is most commonly identified as the first general electronic text corpus is the Brown University corpus (Kucera and Francis 1967), created in the early sixties. Usually referred to as the Brown corpus, this resource measures a million orthographic words (tokens). Twenty years later, a corpus created in Birmingham to assist in the well-known lexicographic COBUILD project (Sinclair 1987) was larger by a factor of ten.[3] The next tenfold increase in corpus size is the 100 million words of the British National Corpus (BNC: Burnard 1995; Leech et al. 1994), also created with lexicographic applications foremost in the mind. The British National Corpus has become a *de facto* standard of a national corpus, a model of sorts for other similar undertakings (cf. Fillmore et al. 1998). It is worth stressing at this point that the complete body of the British National Corpus, and a subset of the Bank of

[3] The Bank of English, which has grown out of the COBUILD corpus, has remained among the largest corpora of English, with the most recent available reports (http://www.titania.bham.ac.uk/docs/about.htm) placing the size at somewhat above 500 million words, although it is to a large extent opportunistic. However, currently at least two other corpora are said to have exceeded 1 billion words of text: The Cambridge International Corpus (http://www.cambridge.org/elt/corpus/what_can_corpus_do.htm) and the Oxford English Corpus (http://www.askoxford.com/oec/mainpage/?view=uk).

English corpus (which has evolved out of the COBUILD corpus) are now searchable through the WWW.[4]

Today, the largest corpora have grown by yet another order of magnitude. The Bank of English, which has evolved out of the COBUILD corpus, has remained among the largest corpora of English, with the most recent available reports[5] placing the size at somewhat above 500 million words, although the corpus is to a large extent an opportunistic one. However, currently at least two other corpora are said to have exceeded 1 billion words of text: The Cambridge International Corpus[6] and the Oxford English Corpus.[7]

Turning to the size of the World Wide Web, the approximate size of the textual resources of the WWW can be extrapolated from the number and average length of documents indexed by search engines. Before it stopped publicizing the number of indexed pages in late August 2005 after the famous 'size war' with Yahoo,[8] the most popular search engine Google[9] claimed the number of pages in its indexes to be over 8 billion, by a very conservative estimate (excluding, e.g., partially indexed pages). Applying to this number the estimation algorithm proposed by Lawrence and Giles (1999) puts a rough estimate of the total (indexed and unindexed) textual resources at five trillion (5,000,000,000,000) word tokens: that is about fifty thousand times the size of the British National Corpus. Of course, such estimates will vary widely depending the assumption of what type of content should be counted; and, such extrapolation is becoming more difficult and less reliable with the increasing reliance of the WWW on content generated on the fly from some type of underlying database (the so-called deep-web, cf. Bergman 2001). This last source of error is likely to lead to underestimation rather than otherwise, and there is no questioning the fact that the size of the WWW is greater by several orders of magnitude compared to traditional corpora.

[4] http://sara.natcorp.ox.ac.uk/lookup.html, http://www.collins.co.uk/Corpus/CorpusSearch.aspx
[5] http://www.titania.bham.ac.uk/docs/about.htm
[6] http://www.cambridge.org/elt/corpus/what_can_corpus_do.htm
[7] http://www.askoxford.com/oec/mainpage/?view=uk
[8] http://www.yahoo.com/
[9] http://google.com

The rate of growth of the English-language part of the internet can be appreciated by looking at the occurrence frequency numbers for specific phrases and comparing them against the British National Corpus, as shown in Table 1. As the numbers indicate, the current size of indexed textual content of the English-language web exceeds by some four orders of magnitude the size of a large general corpus. One should not forget that, as indicated above, the number of pages reported by the search engine would typically be about one order of magnitude smaller than the number pages that are in fact available on the web, because a large proportion of the web content remains unindexed (Lawrence and Giles 1999).

phrase	BNC	WWW				
		autumn 1998	autumn 2001	spring 2003	2 Dec 2004	16 Oct 2006
medical treatment	414	46,064	627,522	1,539,367	1,960,000	11,300,000
prostate cancer	39	40,772	518,393	1,478,366	2,420,000	15,300,000
deep breath	732	54,550	170,921	868,631	1,770,000	6,010,000
acrylic paint	30	7,208	43,181	151,525	225,000	1,350,000
perfect balance	38	9,735	35,494	355,538	498,000	2,370,000
electromagnetic radiation	39	17,297	69,286	258,186	272,000	1,580,000
powerful force	71	17,391	52,710	249,940	326,000	2,000,000
concrete pipe	10	3,360	21,477	43,267	63,700	648,000
upholstery fabric	6	3,157	8,019	82,633	111,000	981,000
vital organ	46	7,371	28,829	35,819	59,200	207,000

Table 1: Frequencies of occurrence of selected English noun phrases in the British National Corpus (BNC), AltaVista[10] (1998, 2001) and AlltheWeb[11] (2003, 2004, 2006). The figures for the BNC and AltaVista refer to the number of individual tokens. AlltheWeb cites the number of pages, so there may be more than a single token of a given phrase on a single page. The 1998-2003 data are taken from Kilgarriff and Grefenstette (2003).

Admittedly, 100 million is a very large number of words. However, the statistical nature of the distribution of lexical items in natural text is such that the large majority of tokens turn out to be forms of the most frequent lexemes, while the number of occurrence of tokens representing the less common words decreases exponentially (as described by Zipf's law, Guiraud 1959; Zipf 1935). Thus, while a corpus of 100 million word tokens is large enough to adequately represent the systematic facts of syntax (though see some reservations further down in the chapter), when it comes to lexical facts, a 100-million-word corpus gives a reasonably accurate picture for at most the 10 thousand most common lexemes. Less frequent

[10] http://www.altavista.com
[11] http://alltheweb.com; I specifically avoid using Google hit counts because of their now infamous instability, on which see e.g. http://aixtal.blogspot.com/2005/03/google-snapshot-of-update.html

items are represented by fewer than 50 occurrences each, which does not provide a strong enough basis for statistically stable generalizations (Kilgarriff 2003; Kilgarriff and Grefenstette 2003).

It is also telling that, as shown by Banko and Brill (2001), the effectiveness of resolving lexical ambiguity grows monotonically with corpus size up to the size of at least one billion words. Now, language engineering applications are not our primary focus, but this empirical fact is suggestive of the wealth of linguistic information that is potentially usable in corpora of different sizes.

The issue of corpus size is also related to the epistemological problem of negative evidence: in principle, the fact that a given form is not present in a corpus cannot be used as deterministic proof that the form is a nonexistent one. This is true of a sample of any size, except when the sample consists of the whole population (but that is, arguably, impossible for the population of utterances or texts of a living language). However, in terms of statistical inference and fuzzy logic, the bigger the corpus, the stronger the basis for claiming the non-existence of a form from its absence in a corpus.

The size range of a language resource needs to be appreciably larger if it is to provide a useful coverage of lexical combinations: idioms, phrases, and collocations. This is so because the textual frequency of the cooccurrence of two or more words is naturally smaller, and often very much smaller, than the occurrence frequency of each of the component elements separately (compare the low frequency figures in the BNC column of Table 1 for noun phrases which are, subjectively speaking, not at all uncommon).

While it is probably a relatively safe assumption that very infrequent lexical items are not the primary interest of a foreign language learner (but may be of interest to a linguist!), the language learner will no doubt want to be able to learn about patterns of semantically-motivated lexical co-occurrence (to avoid using the variously understood word *collocation*).

This is all the more important for the fact that lexical co-occurrence information is hard to find in a dictionary,[12] and indeed it is actually difficult to *represent* lexicographically in a satisfactory manner.[13] In the light of the above, it seems that the issue of corpus size would be quite important to both the linguist and the language learner. This quantitative aspect appears then to score a point in favour of the World Wide Web, when seen in opposition to traditional corpora.

3. Linguistic representativeness and the balancing of corpora

The issue of corpus representativeness cannot be usefully taken up unless one specifies the population (in the statistical sense) which we would expect to be faithfully represented. But, as noted by Sambor, we are dealing here with 'trudność natury ściśle lingwistycznej – nie istnieje mianowicie żaden jednorodny makrotekst jako populacja generalna, wobec której badane teksty można byłoby traktować jako próby z niej wylosowane' [a purely linguistic problem – there exists no uniform macrotext or general population that our set of texts could be treated as a sample of – translation RL] (Sambor 1988: 54-55). So, if we accept the view that there is no agreed standard of comparison, there are serious problems with establishing the criteria for corpus representativeness, and thus the usefulness of the very notion becomes questionable, at least for a general corpus: perhaps specialized corpora or text genres might be more easily dealt with.

In turn, the notion of the balancing of corpora is usually taken to refer to the selection of texts that go into the corpus being done in such a way so as not to favour, or disfavour, any particular text type(s); one could say then that corpus balancing is a weaker criterion, and more easily met than corpus representativeness, although some authors actually use the two terms interchangeably (Fillmore et al. 1998; Smarr and Grow 2002). In practice, the text types

[12] Things might be different for syntactically-motivated co-occurence (or colligation, in the terms of Sinclair 1991), where both smaller corpora and dictionaries appear to be mostly adequate.
[13] Although attempts, more or less successful, have been made, notably the recent *Oxford Dictionary of Collocations* (Lea 2002).

most commonly overrepresented in language corpora are press archives and fiction, while the most severely underrepresented type is probably spontaneous speech. This, of course, is a consequence of the grading of difficulty in acquiring linguistic data of a given type.

The unbalanced quality of the language content of the World Wide Web is among the most often listed drawbacks of this resource when considered from the point of view of linguistic applications. Kilgarriff and Grefenstette appear to make light of the problem, maintaining that: '[t]he Web is not representative of anything else. But neither are other corpora, in any well-understood sense' (2003: 343). I do not think that it is fair to equate the World Wide Web with traditional corpora in this regard, as the very nature of the nonrepresentativeness is different in the two cases. For example, it is obvious that an inordinately high proportion of web-based texts are about various aspects of the web itself, which constitutes an interesting kind of systematic reflexivity. Another important point to note is the clear dominance of a single text genre: the web page. Somewhat less obviously, the World Wide Web exhibits an overrepresentation of texts about high technology. While the above features of the web are in fact inherent qualities of the language of the internet, the reasons behind the lack of balance in traditional corpora are more incidental; they are largely within the control of their creators and can be subject to planning activity. For this reason, it seems that the two cases of textual imbalance need to be distinguished on theoretical grounds.

One new threat to the representativeness of the texts on the World Wide Web is the relatively recent practice of some web content creators flooding their pages with a high number of repeated keywords in a way that is unobtrusive to the human reader, such as through the use of a tiny background-coloured font. This practice is known under the term *web-spamming* (Gyongyi and Garcia-Molina 2005), and it is aimed at boosting the position of the offending pages on results displayed by search engines through artificially exaggerating the frequency of occurrence of words known to be frequently sought. Such manipulation can not only affect the positioning of the page in search results, but also lead to the

misrepresentation of lexical frequency distribution figures. However, designers of search algorithms are defending themselves against the consequences of such deceitful practice by adding mechanisms capable of ignoring such artificially added material. Somewhat ironically, the newest and most insidious type of search-engine spam, the so-called *link farms* (an inevitable consequence of the recent increased reliance of search engines on the analysis of hyperlink patterns – see e.g. Drost and Scheffer 2005), presents less of a threat to the reliability of lexical frequency counts than the spamming of web page content proper.

Lack of balance must be seen as a serious problem for both our categories of users of corpus-like resources: the professional linguist and the foreign language learner. By this criterion, then, (balanced!) traditional corpora should be seen as superior to the Web. However, when it comes to the sensitivity to corpus imbalance, linguists may be in a better position to compensate for the lack of balance, what with their expert metalinguistic and linguistic knowledge, as well as a usual dose of scientific scepticism. On the other hand, one important argument for using corpora is the notorious unreliability of intuition for judging linguistic data, so there are limits to such compensation.

3.1. Noise

Texts that go into a traditional corpus are normally subjected to filtering and cleaning procedures. Moreover, not infrequently they will be texts of high editorial quality to start with. Things look very different when it comes to texts available on the World Wide Web, where the proportion of all kinds of errors and mistakes, including typos, is substantial. However, as soon as you compare the alternative forms, doubts should disappear: 'the Web is a dirty corpus, but expected usage is much more frequent than what might be considered noise' (Kilgarriff and Grefenstette 2003: 342).[14]

[14] The errors themselves may actually be of interest to linguists, EFL teachers and learners.

Just as for the dangers related to the lack of representativeness, the risk of misleading the users would be appreciably higher with language learners than with linguists, as the latter have a higher degree of language awareness.

4. Functionality and access mechanism

Proposals to use universal search engines for accessing the textual resources of the Web, either directly or with various types of additional processing, have been made by a number of authors (Kehoe and Renouf 2002; Kilgarriff 2001; Resnik and Smith 2003; Rundell 2000; Schmied 2006; de Schryver 2002; Smarr and Grow 2002; Volk 2002). According to Kilgarriff and Grefenstette (2003: 344-45), for the working linguist, the most serious drawbacks of using search engines compared to dedicated corpora are as follows: restrictions on the number of hits returned, narrow textual context, awkward ordering of citations in the results lists, impossibility to specify linguistic criteria (such as part of speech) in search queries, and difficulty in searching for all wordforms of a lexeme at the same time (i.e., lack of lemmatization or stemming). I will briefly discuss these problems below, with respect to the needs of the linguist and the foreign language learner.

A typical maximum number of citations returned by a search engine is of the order of a few thousand (Kilgarriff and Grefenstette 2003). It seems that a few thousand items should be more than satisfactory for the language learner. This amount of data should in most cases also satisfy the linguist, unless the goal is to generate mass data for further processing (especially as the limits on automatically run queries may be more stringent).

Narrowing the textual context (= co-text) to (typically) a dozen or so words can indeed be awkwardly restrictive to both categories of users considered here. This is particularly true for issues related to phenomena surfacing at the suprasentential level, such as, for example, discourse-linking adverbs. Still, the interested user can in each case expand the context easily and arbitrarily, by clicking the relevant hypertext link.

When it comes to the ordering of results returned by search engines, most of the popular engines now tend to favour to some extent those pages where search keywords are found in structurally prominent positions, such as titles or document section headings. Preference is also given to those web pages which are the targets of hyperlinks from a large array of other sites. While this seems to be a sensible strategy for the typical users of search engines, likely to produce highly relevant results at the top of the list of results, it is an undesirable feature for the linguist or language learner who are normally after examples of *typical use* of language.[15] In some search engines, there are mechanisms allowing the user to at least partially remedy the above problem. Unfortunately, they are poorly documented and largely unknown. In Google, for example, the prefix *allintext:* or *intext:* can at this time be used in the search query to eliminate the structural position bias, though this particular feature of query syntax is not documented anywhere on the Google help pages. In the current version of the MSN/Live Search engine one can control the weight of the page popularity parameter.[16]

The impossibility to specify in search queries linguistic criteria such as part of speech may be a serious limitation to both the linguist and the advance language learner. One should not forget, though, that this disadvantage is shared by a substantial proportion of traditional corpora as well, as such an option is afforded only by corpora properly annotated for part of speech. A partial workaround for the lack of part-of-speech sensitivity might be to limit the query to a single inflectional form that is unique to a given syntactic category, but this is not always a viable option for a weakly inflected language such as English. Also, one must bear in mind the fact that a given inflectional variant of a lexeme may have specific lexicogrammatical patterns which will then be overrepresented from the point of view of the complete lexeme.

---

[15] Of course, such structural bias might actually be an advantage for a researcher of the language of the web interested in, say, document titles.

[16] Available from late 2004, initially as a test feature at http://beta.search.msn.com, now incorporated as a standard feature of http://search.live.com, the parameter may be set by dragging a graphic slider with the mouse under the Advanced, Results Ranking option, or directly by appending the popularity tag to the search expression, thus *{popl=0...100}*.

Searching for all word forms of a single lexeme is an option available only in morphologically analyzed and annotated corpora. Kilgarriff and Grefenstette (2003: 345) note that search engines do not offer such an option. It is indeed true that lemmatization (or stemming[17]) has never been a mainstream functionality of search engines. The early search engines were rather barebones, and although there was a period around the turn of the century when quite a few search engines did experiment with lemmatization or truncation, later they dropped the functionality, one by one. Presumably, lemmatization did not offer a satisfactory commercial return compared to its increased computational cost. At the time of Kilgarriff and Grefenstette's 2003 analysis, of the major search engines, MSN Search[18] offered lemmatization, but at the time of this writing (November 2006) it no longer does. Lemmatization and truncation (as well as proximity search) is currently offered by the Exalead[19] engine. Since recently, Google appears to be offering some degree of (what it calls) stemming.[20]

All these drawbacks of existing search engines have inspired work aiming to give linguists access to the enormous text resources of the World Wide Web through an interface that is similar to those familiar from concordancers used for corpus searching. Such work can be categorized into two types of projects.

The first option is to try and create from scratch a linguistic search engine, optimized to reflect the needs of linguists. [21] Within this rubric, one could name Adam Kilgarriff's (2003) Linguistic Search Engine, the Search Engine for Applied Linguists (Fletcher 2004) – both at

[17] In the most common usage, the term *lemmatization* refers to the representation of all inflectional wordforms of a lexeme by a citation form, or, in the context of search engines, another inflectional word form of this lexeme. In contrast, *truncation* refers to the use of a simple orthographic substring with a wildcard symbol representing any ending, without a true morphological analysis into lemmas. *Stemming* is often synonymous with *lemmatization*, though it is sometimes used in the same sense as *truncation*.
[18] http://search.msn.com
[19] http://www.exalead.com/search
[20] Google's description of the feature at http://www.google.com/help/basics.html suggests some degree of derivational morphological analysis, so that including the word 'dietary' will also find 'diet'. This feature appears to be selectively active, for the less frequent keywords only.
[21] It is worth noting at this junction that the prototype of chronologically the first search engine, Alta Vista, owes a lot to linguistic insight (Kilgarriff 2003).

present at the proposal stage, as well as the Linguist's Search Engine[22], which has a functioning test version running on a relatively small material of 3.5 million sentences. This last resource allows the user to generate and store customized collections as texts, and supports grammatical parsing including a visualization module that draws structural marker trees. This is also the direction in which the WebCorp project (see below) is apparently heading.[23]

The second option is to provide a layer of pre- and post-processing (what some authors call *wrappers)*, which redirect queries entered by the user to existing search engines (using either the hypertext protocol or the API interface), and then filter and present the results in appropriate format. Projects that work according to this principle may be further divided into those accessible through a web-page interface, and applications requiring installation on the users' computer. The first subset includes the following: WebCorp[24] (Fletcher 2004; Kehoe and Renouf 2002; Morley 2006; Morley et al. 2003; Renouf et al. 2006), WebCONC,[25] WebPhraseCount[26] (Schmied 2006), and Lexware Culler[27] (Dura 2006). The best-known application of the second subset is the KwicFinder[28] (Fletcher 2004).

It appears that at the present time most of the services of the second type above do not offer dramatic improvements over the basic search engine functionality, but they do have one significant disadvantage: inferior speed. Searches take much longer than is the case in search engines. This, incidentally, is also the disadvantage of large traditional corpora. In situations when the user needs to query the resource repeatedly at short intervals, the cumulative delay may become unacceptable and thus make such a resource unusable for practical purposes.

[22] http://lse.umiacs.umd.edu:8080, requires registration.
[23] http://www.webcorp.org.uk/webcorp_linguistic_search_engine.html
[24] http://www.webcorp.org.uk
[25] http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi
[26] http://138166.vserver.de/cgi-bin/wpc/cgi-bin/run.pl
[27] http://82.182.103.45/LexWare/English/demo.html
[28] http://www.kwicfinder.com

Paradoxically, in some cases involving heavily filtered searches, recall can fall way below that of a traditional corpus, as shown convincingly by Lüdeling (in press).

Another aspect of the matter is that applications such as the KwicFinder cannot be utilized on public computers such as libraries, computer labs or internet cafes, since they require installation on the host machine.

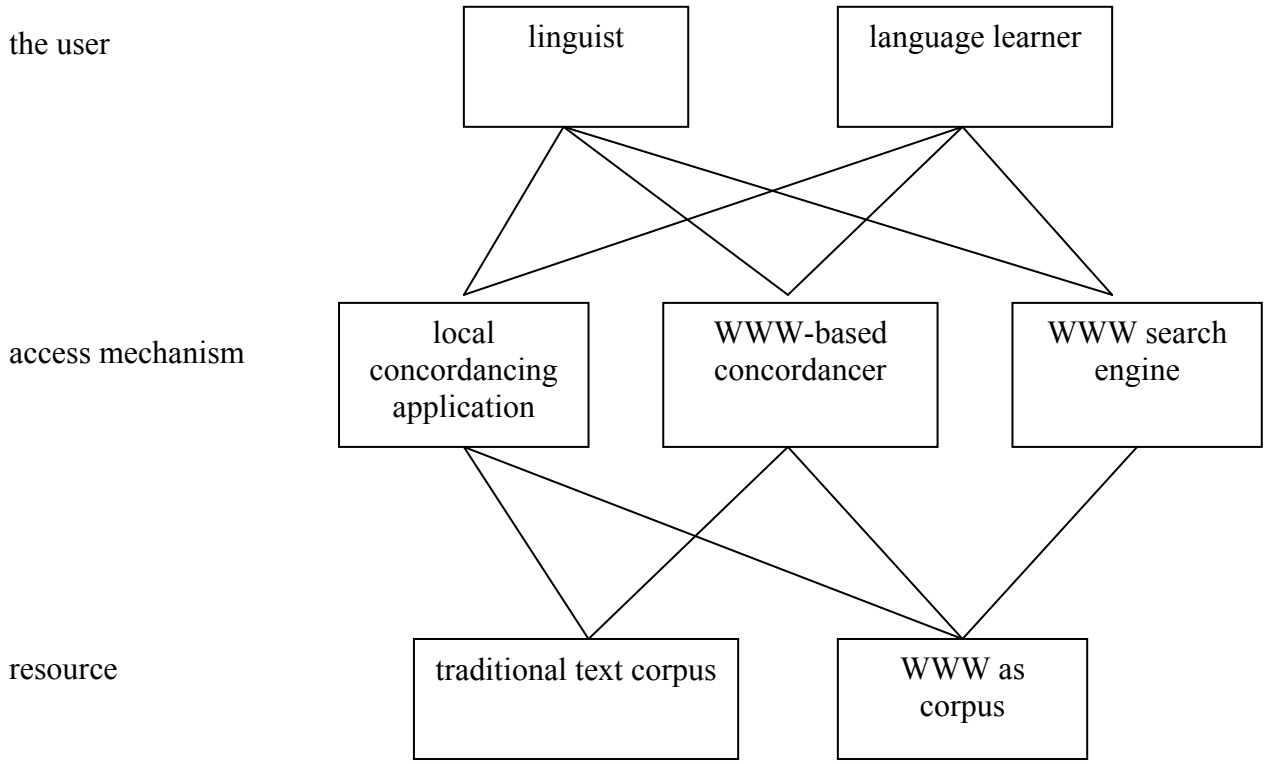
On the other hand, dedicated linguistic search engines sound very promising, especially for the working linguist, although it is probably a little early to attempt a systematic evaluation for what is essentially still at the prototype or development stage.

5. Conclusion

Corpora, being electronic collections of authentic texts, are a valuable source of first-hand language data for the empirically-minded linguist. For the foreign language learner they afford the possibility to verify in an instant, on-line fashion the many working micro-hypotheses regarding language usage against the material representing authentic linguistic behaviour of native speakers of the target language. The World Wide Web can, in this context, be viewed as a unique, dynamic corpus.

The different access configurations to textual resources discussed in the present chapter are presented in Figure 1. For ease of exposition, the Figure ignores more detailed distinctions into further layers of structure within the search application and the search engine. These distinctions are rather important for the information technology expert, but from our perspective they would complicate the already rather complex relations.

Figure 1: Configurations of access to the textual resources of traditional corpora and the World Wide Web by the working linguist and the language learner, divided into three layers.



Based on the above comparison of traditional electronic text corpora and the textual resources of the World Wide Web, it can be concluded that the WWW, despite its noisiness and poor balancing, can be an attractive and useful tool for on-line language reference. Its main virtues lie in the impressive size of the resource, and the speed with which it can be trawled using a general-access search engine. Such a configuration can be helpful in instantly resolving the language learner's immediate lexical problems, as well as serve the linguist in some types of situations. The more sophisticated needs of the working linguist may be better fulfilled by means of traditional corpora or the WWW enhanced with a specialized access interface. Plans to put online dedicated linguistic search engines appear to hold much promise in this regard. At the same time, we should keep in mind that the more sophisticated the tool, the greater its complexity and the skills required of the user. In view of the above trade-off relationship, it seems that for some groups of users, particularly language learners, maximally simplified tools will continue to hold the greatest attraction.

References:

Aston, G. (1997a), 'Enriching the learning environment: Corpora in ELT', in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.), *Teaching and Language Corpora*. London: Longman, pp. 51-64.

Aston, G. (1997b), 'Small and large corpora in language learning', in B. Lewandowska-Tomaszczyk and P.J. Melia (eds.), *International Conference on Practical Applications in Language Corpora, Łódź, Poland, 11-14 April, 1997*. Łódź: Łódź University Press, pp. 51-62.

Banko, M. and Brill, E. (2001), 'Scaling to very very large corpora for natural language disambiguation', *Proceddings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics, Toulouse, 2001*.

Bergman, M.K. (2001), 'The deep web: Surfacing hidden value', *The Journal of Electronic Publishing* 7, (1).

Biber, D., Conrad, S. and Reppen, R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*. *Cambridge Approaches to Linguistics*. Cambridge: Cambridge University Press.

Burnard, L. (1995), *The BNC Reference Manual*. Oxford: Oxford University Computing Service.

Drost, I. and Scheffer, T. (2005), 'Thwarting the nigritude ultramarine: Learning to identify link spam', *Machine Learning: Ecml 2005, Proceedings*, (*Lecture Notes In Artificial Intelligence*), pp. 96-107.

Dura, E. (2006), 'Extracting current language use from the Web', *Poznań Studies in Contemporary Linguistics* 41. 73-85.

Fillmore, C., Ide, N., Jurafsky, D. and Macleod, C. (1998), 'An American National Corpus: A proposal', in A. Rubio, N. Gallardo, R. Castro and A. Tejada (eds.), *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada.

Fletcher, W.H. (2004), 'Facilitating the compilation and dissemination of ad-hoc web corpora', in G. Aston, S. Bernardini and D. Stewart (eds.), *Corpora and Language Learners*, (*Studies in Corpus Linguistics 17*). Amsterdam: John Benjamins, pp. 273–300.

Grefenstette, G. (1999), *The WWW as a resource for example-based MT tasks*. Plenary talk at the ASLIB conference *Translating and the Computer*. London: ASLIB.

Guiraud, P. (1959), *Problemes et méthodes de la statistique linguistique*. Dordrecht: D. Reidel Publishing Company.

Gyongyi, Z. and Garcia-Molina, H. (2005), 'Spam: It's not just for inboxes anymore', *Computer* 38, (10). 28-+.

Johns, T. (1991), 'Should you be persuaded: Two samples of data-driven learning materials', *English Language Research Journal* 4. 1-16.

Kehoe, A. and Renouf, A. (2002), 'WebCorp: Applying the Web to linguistics and linguistics to the Web'. *WWW 2002 Conference*. Honolulu, Hawaii.

Kilgarriff, A. (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 University, 29 March - 2 April 2001*. Lancaster: UCREL, pp. 342-44.

Kilgarriff, A. (2003), 'Linguistic search engine', in K. Simov and P. Osenova (eds.), *Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003), 27 March 2003, held in conjunction with the Corpus Linguistics 2003 conference*, (*University Centre for Computer Corpus Research on Language Technical Papers*). Lancaster: UCREL, Computing Department, Lancaster University, pp. 53-58.

Kilgarriff, A. and Grefenstette, G. (2003), 'Introduction to the special issue on the web as corpus', *Computational Linguistics* 29, (3). 333-48.
Kucera, H. and Francis, W.N. (1967), *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.
Lawrence, S. and Giles, C.L. (1999), 'Accessibility of information on the Web', *Nature* 400. 107–09.
Lea, D. (ed.), (2002), *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.
Leech, G., Garside, R. and Bryant, M. (1994), 'CLAWS4: The tagging of the British National Corpus', *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Kyoto, pp. 622-28.
Lüdeling, A., Evert, S. and Baroni, M. (in press), 'Using web data for linguistic purposes', in M. Hundt, C. Biewer and N. Nesselhauf (eds.), *Corpus linguistics and the Web*, (*Language and Computers - Studies in Practical Linguistics 59*). Amsterdam: Rodopi.
Morley, B. (2006), 'WebCorp: A tool for online linguistic information retrieval and analysis', in A. Renouf and A. Kehoe (eds.), *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, pp. 283-96.
Morley, B., Renouf, A. and Kehoe, A. (2003), 'Linguistic research with XML/RDF-aware WebCorp tool'. *WWW 2003 Conference*. Budapest.
Partington, A. (1998), *Patterns and Meaning: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
Renouf, A., Kehoe, A. and Banerjee, J. (2006), 'The WebCorp Search Engine: A holistic approach to web text search', *Electronic Proceedings of CL2005*. Birmingham: University of Birmingham.
Resnik, P. and Smith, N.A. (2003), 'The Web as a parallel corpus', *Computational Linguistics* 29, (3). 349-80.
Rundell, M. (2000), 'The biggest corpus of all', *Humanising Language Teaching* 2, (3).
Sambor, J. (1988), 'Lingwistyka kwantytatywna - stan badań i perspektywy rozwoju', *Biuletyn PTJ* 41. 47-67.
Schmied, J. (2006), 'New ways of analysing ESL on the WWW with WebCorp and WebPhraseCount', in A. Renouf and A. Kehoe (eds.), *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, pp. 309-24.
de Schryver, G.-M. (2002), 'Web for/as corpus: a perspective for the African languages', *Nordic Journal of African Studies* 11, (2). 266-82.
Sinclair, J. (ed.), (1987), *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London - Glasgow: Collins.
Sinclair, J. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, J. (ed.), (1995), *Collins COBUILD English Language Dictionary*, 2nd edition. London - Glasgow: Collins.
Smarr, J. and Grow, T. (2002), 'GoogleLing: The Web as a linguistic corpus'.
Tribble, C. (1991), 'Concordancing and an EAP writing program', *CAELL Journal* 1, (2). 10-15.
Varantola, K. (2003), 'Translators and disposable corpora', in F. Zanettin, S. Bernardini and D. Stewart (eds.), *Corpora in translator education*. Manchester: St Jerome, pp. 55-70.
Volk, M. (2002), 'Using the web as a corpus for linguistic research', in R. Pajusalu and T. Hennoste (eds.), *Tähendusepüüdja. Catcher of the Meaning. A festschrift for Professor Haldur Õim*. Tartu: University of Tartu, pp. 3-13.
Walter, E. and Harley, A. (2002), 'The role of corpus and collocation tools in practical lexicography', in A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August*

12-17, 2002, Vol.2. Copenhagen: Center for Sprogteknologi, Copenhagen University, pp. 851-57.
Willis, D. (2000), *The Lexical Syllabus*. London: Collins.
Zipf, G.K. (1935), *Psycho-Biology of Languages*: Houghton Mifflin.