WebCorp: providing a renewable data source for corpus linguists

Antoinette Renouf

Research and Development Unit for English Studies
University of Liverpool

Abstract

The many electronic text corpora available nowadays present ever fewer obstacles to a wide range of corpus linguistic study. However, corpora are expensive resources to create and to update, and there remain problems for linguists if they seek access to very large, very recent, or changing language. The World Wide Web, whilst intended as an information source, is an obvious resource for the retrieval of linguistic information, being the largest store of texts in existence, freely-available, covering a range of domains, and constantly added to and updated. Individual linguistic researchers have been trying to retrieve instances of rare or neologistic language use from the web by manipulating existing web search engines. Whilst this strategy is possible, in particular via Google, the output is rather haphazard and not linguist-friendly. The Research and Development Unit for English Studies has been seeking to remedy the situation through the creation of' 'WebCorp', a tool designed to search the Internet and provide on-line tailored access to linguists. A demonstration tool is available at http://www.webcorp.org.uk This paper will report on the research initiative and highlight some of the issues involved.

1. Introduction

A previously unimaginable number and range of electronic text corpora are now available to corpus linguists, from small and sampled collections to very large textual databases. Whilst this wealth of data makes possible many types of corpus-based research, particularly in the formerly rather inaccessible areas of lexis and lexico-grammar, it has inherent limitations. In practical terms, the corpus data and software may not be available without the appropriate computer access, licences, and so on. More fundamental linguistic limitations relate to the size, age and static nature of the corpora, which can preclude certain kinds of linguistic empirical investigation, for instance the study of very rare, new or changing language features.

An alternative source of linguistic information is the web, a publicly available data resource containing a vast and evolving accumulation of texts. Admittedly, this is not constructed or managed with the rigour or for the purposes of a corpus. It is a muddle of multilinguality; it operates a loose definition of 'text' which includes all manner of extraneous matter; text dating is sporadic and linguistically uninterpretable, so that neither the latest coinages nor the elements of language change across time that are undeniably in there are traceable by

means of chronological organisation. Nevertheless, as a renewable resource which in itself costs the linguistic community nothing to create or access, it is worthy of serious consideration.

The web itself is larger than any corpus. Estimates vary, but on the basis of extrapolation from AltaVista figures for sample words, we calculate its size, in terms of the searchable texts, to be currently at over 50 billion words and growing. In addition to size, the web obviously offers range; many specialised textual domains are represented. The problem of identifying these other than by Yahoo or Open Directory will be alleviated in due course, as the URLs become more transparent and the mark-up protocols are tightened up.

Web text is up-to-date. By this is meant not that the web consists exclusively of yesterday's or even today's language use, but that it is not subject to the same delays in creation that dog designed corpus initiatives. Web texts are a combination of old and new, but 'old' by web standards generally means texts from the late nineties and 2000. The web is constantly updated, with several million pages being added every day. Consequently, it provides ever more - and more recent - data, and corresponding opportunities for retrieving fresh findings. In holding texts across time, which contain instances of language change which could be traceable, the web could even meet some of the needs of modem diachronic corpus linguists (Renouf, forthcoming-a).

2. Some Strategies for Accessing the Web as a Linguistic Resource

2.1 Off-line Processing

The web is being targeted as a linguistic resource from various quarters and in different ways. Some linguists, such as KUbler and Foucou (2000), are extracting texts, and meta-texts, which together make up corpora deemed to be representative of something, such as 'general language use', or a technical field. Fairon (1999, 2000) downloads entire newspaper sites for processing. Kilgariff (2001), on the other hand, is currently collecting reference sets of URLs based on the BNC typology, with which a user can create domainspecified corpora by downloading, without copyright infringement.

2.2 Enquiry by Search Engine

Another strategy is to exploit the functionality of web search engines. Standard engines operate by searching the web for factual information containing a specified search term. A small effort of imagination recasts this in corpus linguistic terms as searching the web for contexts containing target word or phrase. A growing number of researchers have been driven (Bergh et al, 1998; Brekka, 1999, 2000), by the absence or insufficiency of evidence in existing corpora for rarer or newer linguistic items and features, to attempt a trawl of the web by this means. Search engines are not, however, designed to accommodate such an approach, and the consequent negotiation entails tedious serial searching and downloading of sometimes individually thin pickings, followed by painstaking manual editing of whole texts. For the word *cull*, featuring heavily in British news from spring 2001 in relation to

foot-and-mouth disease, AltaVista yielded 14,005 returns on Nov. 13t1" 2000, using its Advanced Search Facility. We present the first few, fairly typical contexts, trom which it will be plain that the output, whether or not relevant trom the point of view of topic, is neither linguist-triendly in format, nor rich in relevant instances of usage of the word *cull* itself.

> Wilkes.Barre Scranton Penguins Hockey Club. the Official Website of the AHL A Wilkes Barre/Scranton Penguins are the AAA affiliate of the NHL's Pittsburgh Penguins playing in the AHL.
> URL: http://www.wbspenguins.com/o Related pages 0 Translate Topic: Scranton, Pennsylvania - Sports and Recreation
> Alternative Music, NPR News .KCRW 89.9 FM
> Alternative, Eclectic, World, Pop, Jazz, Electronic, House and Hip Hop music and NPR, PRI, BBC and VOA news. Listen live or on-demand with RealAudio.
> URL: http://www.kcrw.org/o Related pages 0 Translate
> Topic: Live Music Broadcasts
>
> Top Internet News Stories from DataSegment.com
> Top Internet News Stories from DataSegment.com (Top Internet News Stories) URL:
> http://breakingnews.datasegment.com/topjnternet_stories/o Translate
> The Register
> 24 August 2001 Updated: 20:22 GMT. Flirting tops recession-beating US agenda. wireless Tasteless maybe - but shockingly cluefull. 24 August 2001...
> URL: http://www.theregister.co.uk/contentl7/index.htmlo Related pages 0 Translate

Only at AltaVista Return 8 does a potentially relevant context occur:

> BBC News | FOOT AND MOUTH
> HOMEPAGE.| SPORT. | WEATHER. | WORLD SERVICE. | MY BBC. Search BBC News Online. "> You are in: In Depth: Foot and mouth. Front Page World UK UK...
> URL: http://news.bbc.co.uk/hi/english/in_depLuth/default.stm ₀Related pages ₀Translate

As a linguistic search tool, Google is unique in extracting context for search terms, and has built in some refinements which will be shown later, but it retrieves only one instance of a search term from a given web page on which it may actually occur several times. A search engine also covers only a slice of the web. Furthermore, it can retrieve information only for the search terms in its periodically-updated index. Google could not yet trace *Sophiegate*, of April 1st 2001 vintage, in early May.

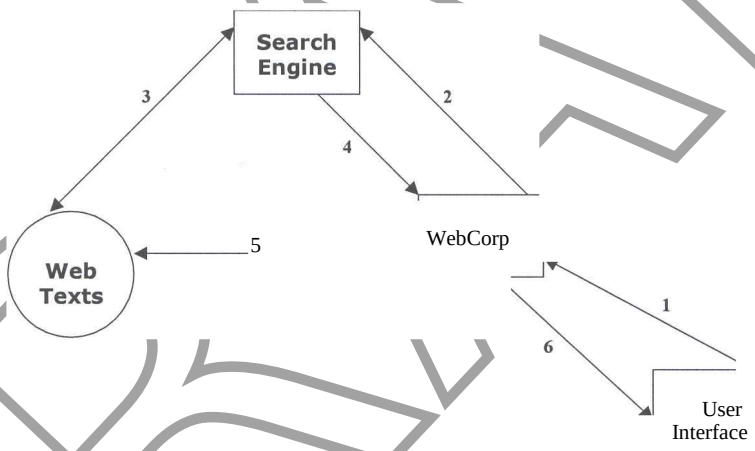
3. A New Linguistic Tool: The WebCorp System

A further leap of imagination reveals the web to be ripe for exploitation by software tailored to find and retrieve contextualised instances of words and phrases. Such information could serve the linguistic community in areas of linguistic, pedagogic, lexicographic and other

endeavour by filling the information gaps left by traditional corpus data. This was the point of departure for the WebCorp project, in the Unit at Liverpool. I intended to set up this project in 1996, but other project priorities meant that it was finally launched in December 2000. The project team consists of Mike Pacey, Andrew Kehoe and Jayeeta Banerjee, software developers; Paul Davies, statistical consultant; and myself, linguist and project originator and manager. Michael Hoey, as PI, is on hand with linguistic advice; Themis Bowcock, as Cl, steers us through web and post-web developments. The UK company Searchengine.com is our industrial collaborator. The WebCorp tool is being developed according to an intensive and ambitious two-year project plan, which has been informed in part by the copious feedback received in response to a simple pre-project prototype software demonstrator that we installed on the web back in May 2000. Among the many unsolicited expressions of enthusiasm for the WebCorp tool, Michael Rundell stated in his paper entitled 'The biggest corpus of all' (Rundell, 2000) that:

"...a major breakthrough is at hand, in the form of a stunning new website that produces real 'concordances'. As with Altavista and others, http://www.webcorp.org.uk/ [Le. WebCorp] searches the entire Internet for your query. But in this case the output is a proper concordance with an amount of surrounding context which the user (that's you) can specify in advance. The results, in other words, look very similar to what you might get from the BNC or COBUILD Direct – but in this case the "source data" is the vast store of text on the entire Internet."

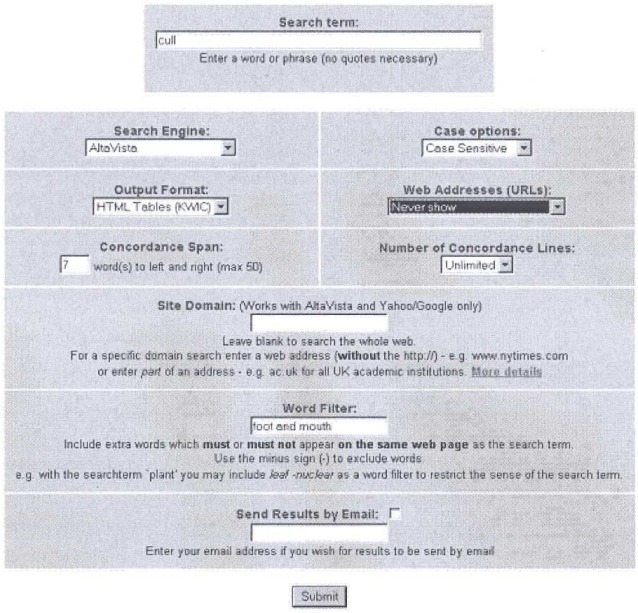
3.1 Diagram of WebCorp System Operation



The WebCorp system operates as follows. The tool currently has six stages, as shown in the graph above. It first interfaces with the user request, which can be a word or a contiguous phrase, converting it into a format acceptable to a selection of search engines. It then piggy-backs on one or other of these that has been specified by the user. Each search engine

follows its own procedure for searching a section of the web for texts containing the specified language item. Once the engine has traced the search term, via its own index, to a candidate text, WebCorp down loads that text temporarily into memory and extracts the appropriate linguistic context, processing and collating it before presenting it to the user.

Graphical User Interface of the WebCorp Tool



In its current form, the Graphical User Interface looks as shown above. It offers various options to the user. The user can submit a keyword consisting of a word or phrase. The user can select one of 5 search engines: currently, Google, Altavista, Northern Light, FAST, and MetaCrawler. Case sensitivity may be specified or not. There is a choice of output format: HTML, HTML KWIC tables, and plain ASCII text. The user may wish to have a display of URLs, or for the sake of readability, omit these and have an unbroken set of KWIC contexts. The number of concordance lines may be specified. For the purposes of search refinement, a particular site domain may be specified in terms of any part of a URL; for example, *.fr* or . *QC. uk.* The concordance span may be set at between 1 – 50 words to the left

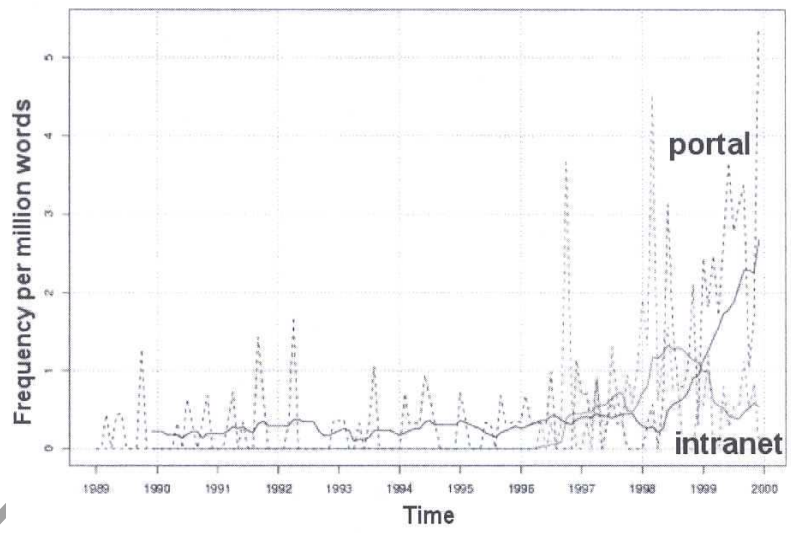
and to the right of the node. Users not wishing to wait for results may receive them by selecting an email option. Further refinements to this interface are in progress.

4. Sample WebCorp Searches

The next section of this paper will present some of the linguistic information which can usefully be retrieved from the web.

4.1 Study of Neologisms

One area of linguistic interest is the search for neologisms and new word uses. One might have noticed in computing journals that a new term, *corporate portal*, appears to be taking on some of the role of the earlier term *intranet.* The APRIL project graph (Pacey et al, forthcoming) confirms a changeover in the frequency with which the two terms *(intranet* and *portal*, since the latter occurs only in *corporate portal)* occur in the Independent newspaper from 1989-1999.



A search using the WebCorp tool with the AltaVista search engine yields 57 occurrences for "corporate portal", which can be extracted in neat, one-line contexts as follows:

1, is using a mobile bizli.com	corporate	portal	to give traveling consultants access
, an employee-only next	corporate	portal	to access intranet applications from
month by delivering a 10,	corporate-portal		platform and related tools capable
1999 TIBCO launches	corporate	portal	building effort, hosting service TIBCO
joins forces with TIBCO on	corporate portal		Yahoo plans to get into

plans to get into the	corporate	portal	business through a partnership with to
developed what we call a	corporate	portal	power this Internet business software on
scalable, widely deployed	corporate	portal	the market. < market-leader goes global!
The	corporate	portal	and e-commerce application
intranet programming	corporate	portal	based on state-of-the demand. movement
Management of your	corporate	portal	is the runaway intranet Seagate Software
The genesis of the	corporate	portal	distributes business space By lan Lynch
Using Domino to build a	corporate	portal	intranet technology by enabling users
News Yahoo and SAP target	corporate	portal	
would spice up Yahoo's	corporate portal		

4.2 Study of Rare Uses

One of the known processes of change in modem English grammar is the shift trom *person who* to *person which* type of construction discussed by Mair (1998). The WebCorp tool can furnish examples of such rare language use, which supplement the few found by Mair in a number of corpus sources, and thus help, as Bergh (1998) says, to 'improve predictive power in determining what is 'real English'.

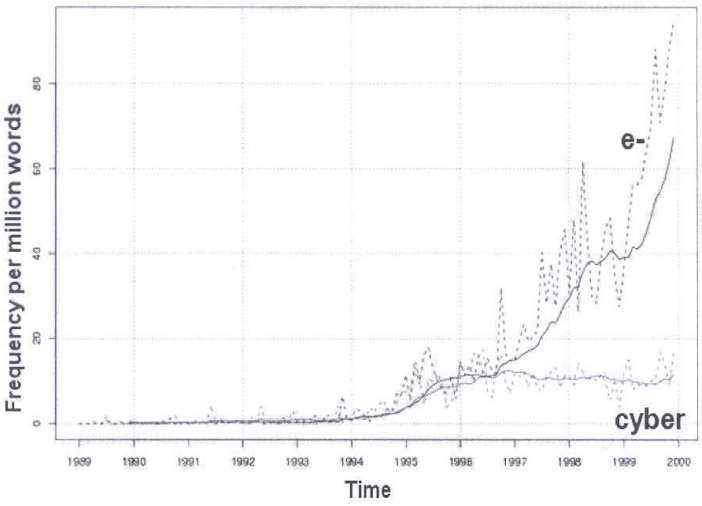
WebCorp used Altavista to generate 88 returns for person which; an extract of which is shown:

on self entries. Here the decide	person	which entered his entry can
on the plan, or by any related	person	which is a disqualifIed person with was
person" means any to make fIrst	person	which under common control
contact. The your fIrst contact.	person	which has written the search ad for has
The Registrar: a registrar is the	person	which written the ad will
46: In Reply toL The	person	which is authorized to enter and posted it
You are the	person	which has to be me guestbook! visits my
same person or by any	person	which page
The	person	which controls or is controlled by, or posted
15 years, or both. A	person	which it has to be
passed by, He saw a	person	which is an organization shall,
by a household or	person	which was blind trom birth.
	person which	generates less than 100 kilos

4.3 Study of Productive Linguistic Features

Another area of development in the English lexicon that one might wish to study is the productivity of some morphemes. However, the chronological, quantified diachronic study that productivity implies to a corpus linguist is not yet possible on the web, due to the lack of textual authorship dating. Nevertheless, hypotheses based on diachronic data sources, such as the chronologically-stored and processed journalistic text collections and the APRIL morphological database at Liverpool; or the FLOB and Frown corpora at Freiburg, can be tested against the larger web resource. The user might note that the *'e-'* splinter, an abbreviation of *electronic*, is gaining status as a full-blown affIx, and is taking over in

popularity from *cyber* and *techno* in creating new formations. The APRIL tool compares *e*and *cyber* frequencies of use in 10 years of Independent text below.



E-words can be searched for with our WebCorp tool. For instance, a sub-set of the *'e'* words is evolving in text, whereby an *E-word* is created not simply by attaching' *e'* to a root form, but by taking a word like *retail,* which has *e* as the vowel in its initial syllable, and clipping the letters preceding the *'e'*. Neologisms arise in this way from semantically transparent and appropriate bases such as *freelancers,* which becomes *freelancers;* and *retail* which becomes *etail,* as shown below:

WebCorp output for search term "etail"

1998 - Web-based retail, or	etail	as the word du jour
Weeks. Close behind are the	etail	enablers, those that make the effort
11.5 million if the	etail	meets targets over next
Can be viewed at www,	etail	.ie home I web sites I brochures has
Providing goods from one supplier	etail	arrangements with several top on your
Your purchases are made through	etail	behalf. Members of will receive
On your behalf. Members of	etail	discounts on the
Normal RRP of products since	etail	h negotiated discounts with all
shopping experience now. ..	etail	Shopping
Personal details. Jump to About	etail	Shopping. The Checkout New Special
Secure Shopping Contacting	etail	Offers for Your Shopping Shopping
Your Shopping Trolley Login to	etail	Search Help

4.4 Study of Unconventional Use

In corpus-based study of language change, one can encounter unexpected usage in otherwise conventional sources. Investigating the continuing regularisation of irregular verbs (e.g. Mair, 1998), for example, one finds in the Independent newspaper several linguistic and metalinguistic uses of *maked, gived* and *catched:*

> It maked sense to police chiefs
> It does not matter to them who maked their laws
> Past tenses as . hited', . maked', . singed' and' goed'.
> Technical papers gived the total number of fuel elements changed
> No sooner had I said this than Guiseppe catched hold of my hand
> The girls bowled, batted, ran and catched as well as most men could

WebCorp would be a resource to turn to for confirmatory or counter evidence. However, alongside conventional usage, the Web is strewn with typographical errors, as well as uninformed, colloquial, provisional and improvised language use of the spontaneous kinds encouraged particularly in chat rooms and news lists; and Web-based text is very often also written by non-native speakers. So one can find instances of almost any odd formation. There is not room in this article to indulge in detailed illustration of the full joys of Web invention; nor is it pertinent. The point is that it means that the Web is not a place to go for reliable confirmation of correct usage unless the user is in a position to evaluate what he/she finds. Such evaluation is usually achievable through native-speaker intuition, but these means are not open to the non-native speaker-learner. There is a need to discover some clues as to the status, bona fide or otherwise, of what is being observed, that can allow the language learner to recognise error and thus make the web a more usable language source.

If we take a questionable form such as *maked*, we find 92 instances yielded by WebCorp via Google on 14th June 2002. Closer observation reveals among these a range of error types, often errors in combination. The purpose here is not to identify the kinds of errors that occur per se, but to see whether there is anything in the environment which can indicate to someone with no recourse to native-speaker judgement that a word is erroneously or conventional spelt and used. Of the 92 instances of *maked*, 54 are native-speaker typographical errors, intended to be the words *marked* or *naked*, as shown in examples 1-4:

1. All mandatory items are maked with an asterisk.(*) (450ccs.)
2. Assessment is by internally maked homework and external examination. (3 occs.)
3. Within a week of starting Accolate, I noticed a maked improvement of my symptoms. (1 occ.)
4. it was considered great fun to eject maked classmates out into the winter (5 occs)

These errors are identifiable to the native-speaker through contextual clues in the form of more, (or less) established collocates (here, respectively: *items, with, asterisk, *; assessment, homework; noticed, improvement; fun, eject* and *winter).* WebCorp currently offers a collocational profile which reveals some of this information (see 6.2).

A further 8 instances appear to be mis-typings of *makes*, created by native speakers out of carelessness or perhaps through a last-minute decision to change tense. The fact that they occur incongruously in text otherwise set in the present is the clue that aids this interpretation. See examples 5-6:

5. If…he **has** only just realised this, it **maked** one wonder if he **has** ever played the game 6. In 'Living on thin air' there **was a prediction that** in the future the tax system **will be** under threat. As companies **become** bigger it **maked** the economies of scale possible

In 2 above, the writer appears to be caught between the conventional rules of reporting, which require past tenses to follow the reporting phrase 'there was a prediction that', and the informal tendency not to sequence reported verbs.

Then there are 11 instances of *maked* which are intended to be *made*. These fall into two categories: 9 that occur in contexts liberally sprinkled with orthographic and grammatical inaccuracies, often on message boards or chat rooms (so identifiable in the URL by such key words as *messages, games, club, connection,* or dubious elements such as *dontlike* or *diy)* - and so perhaps recognisable as errors, as in:

7. The park looked **decidely** scruffy, and stupid things like Submission still having the Virtual Queue lines **maked** the park look like a tacky fun park
8. strumming while Derek (Fish Dick) **maked** his way **throught** the crowd

There are 2 further instances of *maked* for *made* which are isolated errors in otherwise accurate text, and so probably slips of the brain, but correspondingly less identifiable as errors by the non-native speaker. See examples 9-10:

9. Club stalwart Alan Elliott **maked** the trip up to get a day's sailing under his belt
10. One of the fans in one of the PC's started to **maked** a noise. Upon investigation (with the intention of replacing the fan)

There seem to be just 12 instances of non-native speaking errors, all where *maked* is intended to express the past tense of *make*, as in 11-12:

11. Alhambra forests ... **maked it necesary** the application of an special program of control
12. the cinema called "de pipas" ("of pipes") because it was **maked** only to **win** money and **without interest**.

Instances like 11-12 may be recognised by some non-native speakers by the fact that they occur in text authored by people with non-English names, or containing a number of non-native speaking errors as indicated.

The remaining 7 instances of *maked* are either not erroneous, as in the 3 cases of Early English spellings; see 13:

13. Ofauentours bat fel bi dayes, Wher-ofBretouns **maked** her layes."

or they are deliberately wrongly spelt, employed as jokes in 3 cases, as in 14:

14. He 'can not believe you wood get someone who **maked** all those gramar mestakes to WORk for you'

and once in a literary (science-fictional) context, to convey a sense of atmosphere through imitating spoken dialect, in 15:

15. Eye Rock or someplace such, just like as the South'uns **maked** a gather for themselves and clans out at the West

It seems, then, that it is orthographic and lexical clues in the immediate and larger context of a particular word instance that are the learner's best hope of knowing how to judge its accuracy. At the moment, there is no standardised referencing of Web documents such that they are marked as to native-speaker. The current state of Web annotation poses problems for all serious users as to how far the data can be relied on to model acceptable or typical use for pedagogic and lexicographic applications.

Domain specification (see 6.4), whereby message boards and chat rooms are ignored, would help. Country specification, such as *. uk* (see 6.5), might limit the retrieval of non (English)-native-speaking text and word use (though not careless or humorous native-speaking contributions). An automated spell-checker; that is to say a master word index against which items were checked and graded, would be another useful filter, although one man's error can be another man's creative use, and today's error next year's norm. A cumulative collocate bank for the words indexed could rate a Web word for probable misspelling on the basis of its context. It is hoped that the next generation of XML annotation, together with advances in the Semantic Web initiative, will increase the chances of providing some metalinguistic guide as to the status, provenance and thus reliability of the words under investigation.

5. The User

5.1 User Search Habits

As said earlier, the design of the WebCorp System is informed by usage and user comment via the feedback mechanism on the web. Among other things, we note the types of terms that users tend to submit. A sample of these is shown below:

Sample WebCorp Search Terms

abruptity	eskamoteur	a.sc.
albeit	rehabilitation	spinach is bad
diffuse tension	virtual classes	pastoral
token	shockjock	steven bird
sophie-gate	market drive	elder rights
blandity	hear back ITom	hopefulty
milleaux	misfortunate	gargle
the better	tip of my tongue	fixed for
at long last	disinvestment	snafu
"geht sich aus"	coolosity	kicked the bucket
gave a lecture	la-di-da	gobsmacked
cleave	ball park	00*
brockoli	tweenager	holistic
doing my head in	ick	racialism
odiferous	hot-desking	is your call
can reach me	if you don't mind	zine

5.2 User Requirements for Improvement

The features and improvements most commonly requested by users, according to our feedback tool, were in May 2001 - in descending order:

. Increase speed
. Regular Expressions, pattern matching, wildcards
. Full sentence output
. Language selection / detection
. Add more search engines (FAST, Northern Light)
. Ability to customise max. no. of concordances to be returned – useful for slower connections
. Collocation
. Discontinuous phrases – words within certain distance
. Support for double-byte characters (e.g. Chinese)
. More elimination of duplicate results, on same/different pages/sites
. Sorting on left or right words
. *Senseval-style* output format option I
. Wild cards

I *Senseval* is an initiative established in the US as a means of evaluating competing software tools for their efficacity in identifying sense relations by various means.

6. Improvements in WebCorp Functionality

Taking into account the user feedback above, but also following our own planned programme of development, we have moved through successive versions of the WebCorp tool. Progress is incremental but swift. Since the prototype tool was first reported on at the Corpus Linguistics 2001 conference in Lancaster, two new versions have emerged, the first offering improvements including smaller font and compact presentation for concordance lines, numbered concordance lines, and HTML-centred keywords.

6.1 Formatting

By way of illustration, I show below the application of one particular set of format options, producing a to-word context, HTML-formatted concordance extract with keyword emboldened and line numbering for convenience of reference. This was retrieved via the Northern Light search engine at midday on May 2nd, 2001, (at a time when the neologism was still not accessible in the indexes of either Alta Vista or Google).

Sample WebCorp output for the recent neologism, 'Sophiegate':

I. isn't likely to end" Sophiegate " soon. Word is, the newspaper
2. called R-JH. Thanks to Sophiegate , she's stepped down and
3. the sharp end of the Sophiegate skewer. Tony Blair put on
4 Britain's closet republicans. Sophiegate is a huge blow to
5. monarchy faces tough choices over" Sophiegate " tapes. Apr 06 2001 17
6. interest between the two. The" Sophiegate " affair will also be a
7. to say that the recent" Sophiegate " scandal involving Sophie Rhys-Jones
8. lucrative deal. The so-called" Sophiegate " scandal led many newspaper edit 9. Pak-origin scribe set up' Sophiegate '. Pope commemorates Good Friday 10. isn't likely to end" Sophiegate " soon. Word is, the newspaper
11. column Marketing & PRIPress & publishing' Sophiegate ': what the papers s 12.2001. Mark Lawson on the Sophiegate. 31 Mar 2001. Mark Lawson

Thus, in this particular case, WebCorp was able to extract up-to-the-month results for a vogue formation.

6.2 Collocation

The most recent publically-available version (4.7) of the WebCorp tool incorporates type/token counts for web pages, improvements in speed of search and retrieval, and contiguous collocational statistics. Simple collocational information, based on a word span of 4 words to the right and left of the keyword, is shown below, taking *rage* as the search term.

Extract of Collocational Profile for "rage" (excluding stopwords)

Word	Total	L4	L3	L2	L1	R1	R2	R3	R4
air	43				41		1		1
UK	22	8	1	5			1	1	6
computer	6				6				
98	6		1					4	1
rage	6	1	1	1			1	1	1
Britain	6	3					2		1
new	5	1		2					2
International	5						3	2	
Aggression	4	3							1
Internet	4					2	2		
Air	4				4				
Transport	4							2	2
introduced	4		2					2	
99	4		1					3	
incidents	4	1				3			
links	4						2	2	
Sep	4						2		2
so-called	3			3					
Spotlight	3		3						
BA	3	1	2						

Key Phrases: air rage computer rage Air rage

6.3 Precision through Additional Search Terms

Precision in retrieval of linguistic information is required, as it is for information retrieval. That is to say, the user wants to see relevant and only relevant search results; particularly from a heterogeneous environment like the web, which will often swamp the user without some kind of filtering. Google is currently the only search engine which refines search by means of additional search terms. Taking the word *cull*, I assume that the user wishes to test the hypothesis that the word has changed in meaning since the foot and mouth epidemic began this spring; that it no longer means 'strengthen a herd by removal and slaughter of the weaker specimens', but has simply become a euphemism for 'kill'. First using Google's Advanced Search, I specify that the contexts returned must be 'in English', last updated in the 'past 3 months', and must contain the phrase 'foot and mouth' somewhere in the text. I

am presented with the following output on Nov. 13th, 2001, extracted from a total of 4,150 returns:

BBC News I FOOT AND MOUTH
... Farm vaccine report launched Finnie presses for meat exports Foot-and-mouth clean-up complete Cull delay 'worsened epidemic' Legal threat over pyre clean-up ...
Description: Ongoing collection of news articles, reports, forums, audio and video. From BBC News. UK.
Category: Societv> Issues> Animal Welfare> Farming> News and Media
news.bbc.co.uk/hi/english/in_depth/uk/200I/foot_and_mouth/ default.stm - 44k - Cached - Similar ~
Guardian Unlimited I Special reports I Special report: foot ...
... 09.10.01 Foot and mouth inquiry told of 'needless killing'.
04.10.01 Swifter cull 'would have curbed foot and mouth'. ...
Description: Ongoing collection of news, commentary, audio, graphics and interactive guides to the outbreak.
www.guardian.co.uk/footandmouth/O.7368.441391.00.html- 66k - Cached - Similar pages
Foot it around and Mouth Off in Edinburl!h
... If you want to get involved in debate around the Foot and Mouth cull then follow this link. site m~. .
www.mouthoff.org.uk/ - 3k - Cached - Similar pages
Foot and Mouth Disease (FM») site presented by Cybersavvy UK
... a report which states that the cull came too late and was scientifically... using a pneumonia cure to stop foot and mouth reaching his livestock. Although he ...
Description: For Cumbria and the Yorkshire Dales. Latest figures, headlines, links to FMD resources, commentary,...
Category: Society> Issues> Animal Welfare> Farming> News and Media www.webpr.co.uk/fmd/ - 47k - Cached - Similar pages
Yorkshire Dales - foot and mouth - news from Daelnet
... FIGURES showing that hundreds of thousands of cattle slaughtered in the foot and mouth cull did not in fact have the disease threw a massive bombshell into the ...
www.daelnet.co.uk/news/foot_and_mouth/foot_and_mouth_lI0501.cfm - 18k - Cached - Similar pages
CIWF Press Releases 2001
... 25th April 2001, FOOT AND MOUTH SPREAD THROUGH GOVERNMENT FIASCO. 23rd April 2001, MASS CULL BRANDED FUTILE AS FOOT AND MOUTH THREATENS DEER. ...
www.ciwf.co.uk/PRs/2001/2001.htm - 13k - Cached - Similar pages
Foot in Mouth
... Articles on Page 2 Foot and Mouth Disease Like lemmings our ...
www.silentmajority.co.uk/FootInMouth/ - IOlk - Cached - Similar pages

It will be noted that this is not easy to read, and that from the linguistic point of view, several contexts have been truncated. In contrast, the application of the 'Additional Filter' option for the WebCorp tool, using Altavista, produces 156 concordance lines on Nov. 13th, 2001, of which an extract (with contexts which could be specified as longer, and with hypertext links to the original text) is shown below:

Britain to update EU over foot-and-mouth against	cull British farm minister Nick Brown
foot and mouth disease, involving the set your	cull of up to 100,000 animals which
edition UK to relax foot-and-mouth to be put down	cull = LONDON, England (CNN) -
after surviving a	cull on a farm in Devon in the
Massive foot-and-mouth	cull begins. Authorities will look to bu
LONDON, England -- A mass nationwide Horse	cull of livestock is being carried out in
Owners FMD Information - stop the vets are	cull Send a link I Link to us BBC
preparing for a widespread	cull of wild animals in the latest de
birds. Also at risk of a possible	cull are herds of deer spotted gathered ago, had
been contained without a major	cull but accepted that the latest cases slaughtered but
the possibility of a wildlife	cull raised fears of a considerably farmland?"
Attempts to further extend any	cull would lead to logistical problems
was concerned that experts involved in the groups	cull could spread the disease on their disturb settled
of animals during a	cull and send them further afield. Colin the current cull
evidence would justify a major	of wildlife. It they were to
to extend nationwide the pre-emptive	cull of healthy animals within two miles the epidemic
23 Mar 011 Wales Mass	cull begins on Welsh border 23 Mar 01
butterfly in danger from foot and mouth	cull One of Britain's rarest butterflies be threatened cull
by the foot and mouth	The Marsh Fritillary's last stronghold
own livestock of any type. The kill/ Foot-and-mouth	cull policy now in place effects us all
	cull challenged -- With the acrid smoke

6.4 Precision through Domain Specification

Another recent WebCorp refinement is domain specification. Web URLs are not yet transparent, and the only alternative offered by search engines is the indexed information provided by YAHOO and Open Directory, which have pre-indexed down loads of the web. Using WebCorp, however, it is possible to restrict web search by specifying a part or all of a site domain (e.g. *.uk* or *fr).* A linguistic question might concern the use of EU terms as part of the globalisation process in languages. If the user wishes to observe the extent to which the French-originated term *acquis communautaire* is being used in English (Renouf, forthcoming-b), it is possible to do so by restricting the URL search to texts within the UK. An extract of the 127 results produced by WebCorp using AltaVista is shown here:

and legislation consistent with the	acquis communautaire Some of the candidates
ouropean legislation, the so-called	acquis communautaire The only negotiations will European
Union forces too much	acquis communautaire on them. That is the
with some parts of the	acquis communautaire was one on which almost
of European legislation, the so-called the	acquis communautaire The only negotiations will
31 chapters of the	acquis communautaire and has closed II chapters
qualified majority. The	acquis communautaire means that the arrangeme
and the implementation of the	acquis communautaire the faster they can join
it difficult to implement the	acquis communautaire However, I am extremely
meeting the challenges of the irrevocably	acquis communautaire but that it also concerns
binding. The principle of' with the ECJ and	acquis communautaire' means that once any legis
the	acquis communautaire would be remedied. The occupied
field and of the	acquis communautaire are unchallengeable? Mr. completely
right. The concept of	acquis communautaire is very important in this
This is part of the	acquis communautaire which is binding on all

Language identification will improve search refinement, whilst foreign language handling will speed up response, possibly also effected through local (foreign) site location.

6.5 Precision through Language Specification

A further refinement in linguistic search is the specification of a particular language for the source text. The WebCorp tool will soon be refined, in conjunction with our collaborating search engine company, to handle different languages. Meanwhile, the means available for restricting language is simple-minded but effective: it is to specify the particular section of the URL which designates country. Thus, to find instances of *cyberterroriste* in French text, one would specify *'fr'* as the 'domain' option. Using Altavista, this generates 8 results, shown below:

WebCorp output for search term "cyberterroriste"

moigner d'une nouvelle menace d'ordre It	cyberterroriste	It On parle plus de d
quelle personne peut devenir un	cyberterroriste	Detruire des donnees
de l'internautes un	cyberterroriste	
Internet est devenu pour le	cyberterroriste	un outil de plus en
mail-bombing qui permettent de s'improviser	cyberterroriste	au moindre risque. On le
veritable demon. Un	cyberterroriste	un anarchiste. A cote de
messages electroniques en provenance d'un et	cyberterroriste	M. Schmidt - Dui. Un informatique
le qualificatif de	cyberterroriste	qui colle au technoheros

7. Future Plans

Our future plans for improvements to the WebCorp system have been identified above. Three additional areas of development are in operation behind the scenes.

7.1 Grid involvement

The first involves the careful monitoring of next-generation Internet activities. The web explosion will lead to its being superseded by the Grid (Foster & Kesselmann, 1999). 'Grid' is not an acronym but a metaphor for the next stage of Internet and Web organisation, whereby electronic data retrieval and processing activities are conceptualised as basic utilities for society, by analogy with gas or electricity, on a global scale. 'Grid' designates the philosophy behind the provision of vastly increased computing resources, entailing such measures as distributed and shared computing processes. By about 2005, this or a similar resource which will in turn usher in a new generation of electronic facilities - hardware, middleware and new distributed ways of storing and accessing text, which will probably involve text being accessed via the replacement Internet but not sitting directly on it. WebCorp is well placed to develop in step with Grid initiatives at Liverpool.

7.2 Internationalisation

The second stage of development involves the internationalisation of WebCorp in collaboration with colleagues abroad. By 'internationalisation' is meant the introduction of measures to allow the identification and handling of other languages on and via the Web.

7.3 Standardisation

The third line of development involves the standardisation of Web text markup, with particular reference to the dating of text. The current dating mechanism is unreliable, uninterpretable and unuseful in linguistic terms. Our experience shows that just over half of web servers return a 'Last-modified' header, but this fails to indicate whether the date, if given, indicates date of authorship, date of extensive editing and updating, date of complete rewriting, or simply date when minor typographical error was removed. The W3C has proposed the 'Resource Description Framework' (RDF) as a meta-standard, one feature of which is intended to require the page author to specify 'a date associated with an event in the life cycle of the resource'. Among the qualifiers are 'Created', and 'Modified,2. These would be valuable sources of information both in modem diachronic corpus study, and in the study of text editions and versions. We are actively supporting this standardisation initiative (Renouf & Kehoe, forthcoming).

8. Concluding Remarks

This paper has discussed the need of the linguistic community for access to a large-scale, renewable source of information about recent and current language use. It has demonstrated that the web, when accessed by WebCorp, offers linguistic evidence that is not supplied by existing text corpora, or which supplements meagre evidence for rarer or older aspects of language use. A basic system functionality is relatively simple to achieve; the real challenges for WebCorp lie in developing a closer understanding ofthe web's structure and content, and in devising ways of compensating for the current limitations of search engines in order to produce a maximally efficient, informative and user-friendly tool.

9. References

Bergh, G./A. Seppaenen / J. Trotta (1998), "Language Corpora and the Internet: A joint linguistic resource", in Renouf, AJ. ed *Explorations in Corpus Linguistics* (1998) Amsterdam: Rodopi, 41-56.

Brekke, Magnar (1999), 'When "Empiry" Strikes Back: A Corporal Confrontation', Norwegian School of Economics, Norway.

Brekke, Magnar (2000), 'From BNC to the Cybercorpus: A Quantum Leap into Chaos?', in: Kirk (ed).. *Corpora Galore* (2000) Amsterdam: Rodopi.

Fairon Cedrick, "Parsing a Web site as a corpus", in: Fairon, C. ed. (1998-1999), *Analyse lexicale et syntaxique: Le systeme INTEX*, Lingvisticae Investigationes Tome XXII (Volume special), Amsterdam/Philadelphia: John Benjamins Publishing, 450 p.

Fairon C./B. Courtois (2000), "Extension de la couverture lexicale des

2 http://dublincore.orgl documents/ dcmes-qualifiers/

dictionnaires electroniques du LADL a I'aide de GlossaNet", *in Actes du Colloque JADT 2000* : *5e Journees Internationales d'Analyse Statistique des Donnees Textuelles, Lausanne.*

Kilgarriff, Adam (2001). 'Web as corpus', *in Proceedings of the Corpus Linguistics 2001 Conference,* Rayson, Paul, Wilson, Andrew, McEnery, Tony, Hardie, & Khoja (eds) UCREL, 2001. pp.342-344

Kilbler, Natalie & Pierre-Yves Foucou (2000). 'A Web-based Environment for Teaching Technical English', in *Rethinking Language Pedagogy: papersfram the third international conference on language and teaching,* Bumard, Lou and Tony McEnery (eds). Peter Lang GmbH, Frankfurt am Main.

Mair, Christian (1998). Last of the old, or fIrst of the new?', in Renouf, A.J. ed. *Explorations in Corpus Linguistics* (1998) Amsterdam: Rodopi

Renouf, Antoinette (in press-a). 'The Time Dimension in Modem English Corpus Linguistics', in *Proceedings of the TALC 2000 Conference,* Univ. of Graz (2000). ed. Kettemann, Bemhard et aI, Amsterdam: Rodopi.

Renouf, Antoinette (2002). 'Shall We Hors-d'oeuvres? The Use and Abuse of Gallicisms in English', in *Syntaxe, Lexique et Lexique-Grammaire. Volume dedie a Maurice Gross,* eds. Laporte, Eric, Christian Leclere, Mieille Piot & Max Silberztein (eds.) (2000). Lingvisticae Investigationes Supplementa 24, AmsterdamJPhiladelphia: John Benjamins Publishing Co. pp. 523-543

Renouf, Antoinette & Andrew Kehoe (2002). 'WebCorp: Applying the Web to Linguistics and Linguistics to the Web', in *Proceedings of 1 I'h International World Wide Web Conference;* Honolulu, Hawaii, 7-11 May 2002.

Rundell, Michael (2000). 'The biggest corpus of all' in *Humanising Language Teaching,* Year 2; Issue 3; May 2000 (http://www.hltmag.co.uk/mayOO/idea.htm)

10. Acknowledgement

I acknowledge with thanks the funding of the WebCorp project by the EPSRC, and the computational ingenuity of Mike Pacey and Andrew Kehoe, who have developed WebCorp software hitherto.