WebCorp: an integrated system for web text search

Antoinette Renouf, Andrew Kehoe and Jay Banerjee

Research and Development Unit for English Studies
University of Central England in Birmingham

Abstract

The web has unique potential among corpora to yield large-volume data on up-to-date language use, obvious shortcomings notwithstanding. Since 1998, we have been developing a tool, ***WebCorp,*** *to allow corpus linguists to retrieve raw and analysed linguistic output from the web. Based on internal trials and user feedback gleaned from our site (*http://www.webcorp.org.uk/*), we have established a working system which supports thousands of regular users world-wide. Many of the problems associated with the nature of web text have been accommodated, but problems remain, some due to the non-implementation of standards on the Internet, and others to reliance on commercial search engines, which mediation slows up average* ***WebCorp*** *response time and places constraints on linguistic search. To improve* ***WebCorp*** *performance, we are in the process of creating a tailored search engine, an infrastructure in which* ***WebCorp*** *will play an integral and enhanced role.*

1. Introduction

The Research Unit is a multi-disciplinary team of linguists, software engineers and statisticians which works to understand and describe language in use, and to apply this knowledge. The language in question has primarily been English, and the applications have primarily been in the fields of information extraction, retrieval and management, but we are also mindful of the needs of linguistic researchers, language teachers and learners, both in English and in other languages.

We regard language is a changing phenomenon, and we thus began early on to build systems to accumulate and process journalistic text chronologically, to complement existing finite, synchronic corpora. When web text emerged in the nineties, we had been analysing evolving, particularly neologistic, language use in very large textual databases for almost a decade. We were thus well placed to appreciate the advantage of web-based text over the increasingly historical entities which stand as representatives of ‘current English’ – web text would allow the fine-tuning of the picture of what is current usage, providing access to aspects and domains of language which were missing from corpora. Web text presented a serendipitous opportunity, and its many well-rehearsed shortcomings were outweighed by the advantages it offered of access to free, plentiful, up-dated and up-to-date data.

2. Current WebCorp architecture

The WebCorp project was an experiment to see whether we could develop a system to extract linguistic data from web text efficiently and present a quality of raw and analysed linguistic output that was similar to that derived from finite corpora and which met users' expressed needs. In 1998, we placed a simple prototype web search feedback tool on our website, which requested and received user impressions and requirements. By 2000, when funding allowed full-scale system development to commence, we already had a good idea of the functionality we were interested in providing. The basic tool was expanded to provide a range of functions, within the limits imposed by our dependence on commercial search engines and the processing capacity of our servers. WebCorp architecture as it currently stands is represented in the diagram in Figure 1, which also explains the search and analysis routine.

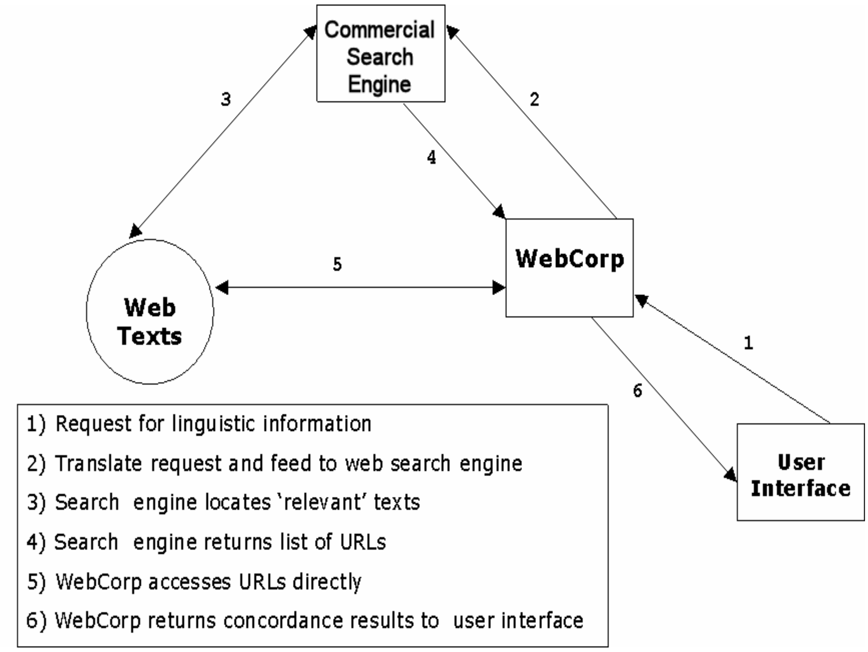
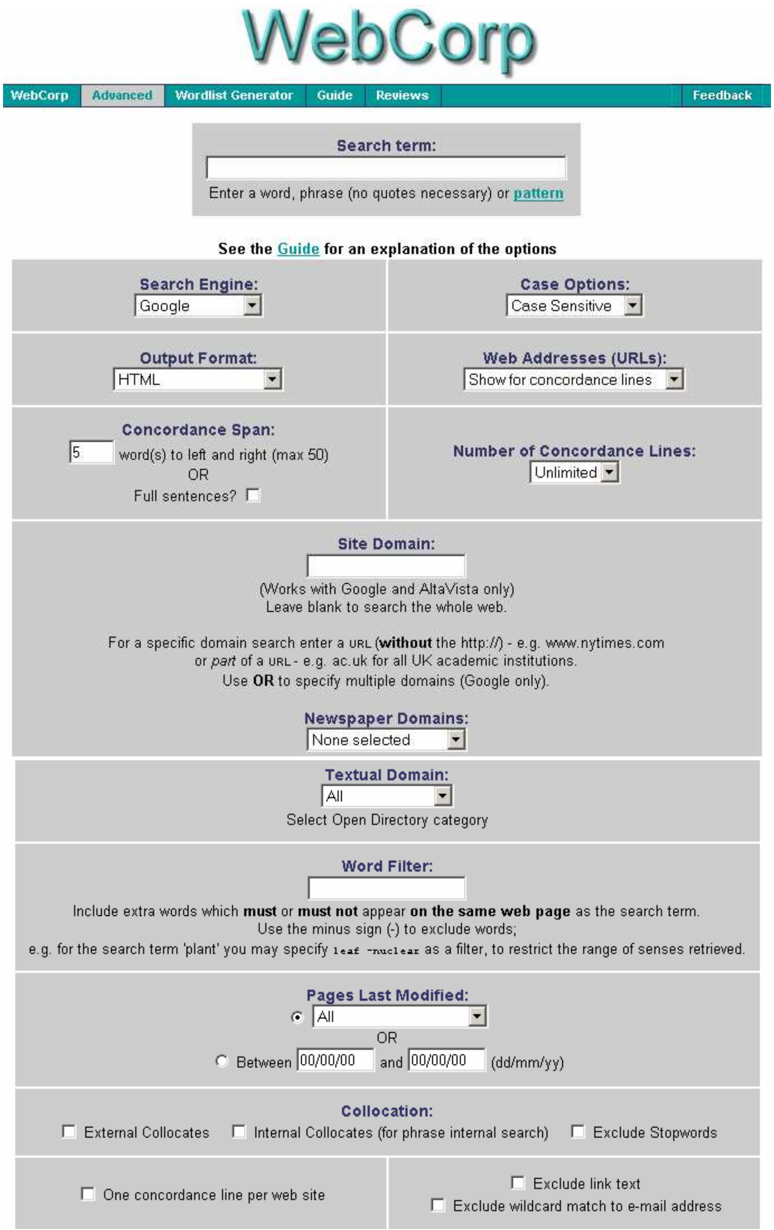

Commercial Search Engine
3
2
4
WebCorp
5
Web Texts
1
6
User Interface
1) Request for linguistic information
2) Translate request and feed to web search engine
3) Search engine locates 'relevant' texts
4) Search engine returns list of URLs
5) WebCorp accesses URLs directly
6) WebCorp returns concordance results to user interface

Figure 1: Diagram of current WebCorp architecture



Search term:

Enter a word, phrase (no quotes necessary) or pattern

See the Guide for an explanation of the options

Search Engine: Google

Case Options: Case Sensitive

Output Format: HTML

Web Addresses (URLs): Show for concordance lines

Concordance Span: 5 word(s) to left and right (max 50)
OR
Full sentences? □

Number of Concordance Lines: Unlimited

Site Domain:

(Works with Google and AltaVista only)
Leave blank to search the whole web.

For a specific domain search enter a URL (**without** the http://) - e.g. www.nytimes.com
or *part* of a URL - e.g. ac.uk for all UK academic institutions.
Use **OR** to specify multiple domains (Google only).

Newspaper Domains: None selected

Textual Domain: All
Select Open Directory category

Word Filter:

Include extra words which **must** or **must not** appear **on the same web page** as the search term.
Use the minus sign (-) to exclude words;
e.g. for the search term 'plant' you may specify `leaf -nuclear` as a filter, to restrict the range of senses retrieved.

Pages Last Modified:
● All
OR
○ Between 00/00/00 and 00/00/00 (dd/mm/yy)

Collocation:
□ External Collocates □ Internal Collocates (for phrase internal search) □ Exclude Stopwords

□ One concordance line per web site

□ Exclude link text
□ Exclude wildcard match to e-mail address

Figure 2: WebCorp user interface: http://rdues.uce.ac.uk/wcadvanced.html

The WebCorp user interface (http://rdues.uce.ac.uk/wcadvanced.html) is shown in Figure 2. As indicated, WebCorp finds words, phrases and discontinuous patterns through word and wildcard search. It currently offers the following series of options for the filtering of information:

- a choice of 5 search engines (of which Google is the most-used)
- upper and lower case distinction
- site domain to help to specify language variety – can be a specific site (e.g. *news.bbc.co.uk*) or a Top Level Domain (e.g. *.uk*)
- a choice of 4 newspaper site groups: UK broadsheet, UK tabloid, French news, US news
- a choice of textual domain based on the Open Directory categorisation, to control language register and probable topic range
- selection of data-subset according to last date of modification
- restriction of the number of instances of an item to one per site, to avoid domination and skewing of results by one author or source
- exclusion of hyperlink text, email addresses and other distractors
- use of a word filter, to improve recall or precision in research results, by allowing or suppressing particular words occurring in the same text as the main search term.

The basic output format takes the form of concordance lines, in each of which the key term is a one-click link back to its full text of origin. The interface also allows the specification of output format in relation to:

- mark-up and layout (HTML or plain text, with or without KWIC layout)
- web addresses (URLs); whether these should be shown in every or any case
- concordance span, in numbers of words to left and right, or as sentence output
- total number of concordance lines

The current functionality of the interface will be illustrated in more details in sections 3 and 4.

3. Types of linguistic information currently retrievable by WebCorp

WebCorp yields large amounts of information about current language use to supplement what is in conventional corpora, but it also opens a window on text domains and types which are not available in corpora, including those which have evolved through its very existence, such as chat room talk. For linguists and language teachers, what WebCorp is uniquely able to provide includes neologisms and coinages; newly-vogueish terms; rare or possibly obsolete terms; rare or possibly obsolete constructions; and phrasal variability and creativity, and we shall demonstrate this facility with a few examples.

3.1 New coinages

A coinage which emerged in web-based newspaper text in July 2005 but which will not to be encountered in designed corpora for some time is the term *deferred success*. This item of political correctness was coined by a UK teaching union official, to replace the word *fail* as a verdict on children's school work. An extract of the linguistic information derivable from web text is presented in Figure 3, which clearly shows the usage patterns and meaning of the word, and indicates that whilst it is too early to see the routine creative inflection and modification of the basic lexeme that would reveal its assimilation into the language, it is already being used humorously (see lines 3-5), and applied in other topical contexts (6-7).

1. The word ‘fail’ should be deleted from the school vocabulary and replaced with the term ‘**deferred success**’, according to a group of teachers.
2. Ms Beattie had in mind when proposing to a teachers' conference that the word "fail" be jettisoned from the educational lexicon, and replaced with "**deferred success**".
3. The phrase ‘failure is not an option’ will be amended to ‘**deferred success** is one of many possibilities on the table’.
4. “When you apply for university they are hardly going to say, ‘Well you have had some **deferred success** so we’ll let you in’.
5. ‘Don't call it failure, call it **deferred success**’, as the bishop said to the actress.
6. A bombing mission has ended in failure or, as politically correct teachers are now being urged to say, **deferred success**.
7. The measure prohibits journalists from describing the situation in Iraq as a ‘failure’ and orders them to replace it with the term ‘**deferred success**’.

Figure 3: results for search term [*deferred success*], filter: UK news

3.2 Rare or obsolete language

Alternatively, the research question may centre on a rare or possibly obsolete item of vocabulary which is not found in existing corpora, and for which confirmation as to its status is sought. An example is the traditional UK colour term *bottle-green*, that seems to have been replaced by such fashion terms as *emerald.* WebCorp nevertheless yields some instances, which are shown in Figure 4. This is useful stuff for the linguist, in that it indicates firstly that the term is not totally obsolete, but only rare, and secondly, that it is used in restricted contexts. It is cited metalinguistically as mention rather than use, in an American online dictionary (1); quoted from 19th century writers, Dickens (2) and Washington (3); used in the scientific context of icebergs (4); and used in reference to gem-stones (5), and to school uniform colours in a colonial context (6). All these instances could be said to reflect use that is either anachronistic, non-UK, non native-speaking English, or semi-technical.

1. **bottle-green** [a] 1) of a dark to moderate grayish green color (thefreedictionary.com)
2. He had a long wide-skirted bottle-green coat on, and a **bottle-green** pair of trousers (*Little Dorrit*, Dickens)
3. bows and arrows…tipped with stone of a **bottle-green** color (*Astoria or Anecdotes of an enterprise beyond the Rocky Mountains*, Washington)
4. The **bottle-green** icebergs of antarctica Antarctic icebergs (Science Frontiers ONLINE No. 87: May-Jun 1993)
5. all **bottle-green** Tourmalines came almost exclusively from Brazil (International Colored Gemstone Association)
6. Hair that touches the collar should be tied up with **bottle-green** hair accessories (Camps Bay Primary School code for school uniform, Zambia)

Figure 4: results for search term [*bottle-green*]

3.3 Phrasal creativity

The phrasal variability and creativity which can be investigated with the use of WebCorp is illustrated with reference to the Chaucerian aphorism *time and tide wait for no man*. This conventional and established idiom can be searched for in its canonical form, but if the linguist wishes to test whether, like all so-called ‘frozen expressions’, it is in fact modified in use, WebCorp supports this activity. The string may simply be submitted with various key words suppressed. Thus, in Figure 5, we see the output of variants forced by the use of the word filter option to suppress the word *tide* in the output.

1. Clear law criminalising identity theft should be introduced as soon as possible. Time, and **cybercrime**, wait for no man.
2. Parliament received a powerful and embarrassing reminder that time and **tights** wait for no man.
3. But time and **semantics** wait for no man and a new volume is deemed necessary

Figure 5: results for search pattern [*wait for no man*]; collocate *tide* suppressed

What Figure 5 reveals, among several other interesting facts about phrasal creativity in general, is that a convention of creative modification is for rhyme, assonance or other phonological devices to play a role in the substitution, as in line 1, where *cybercrime* rhymes with *time*, and in 2, where *tights* assonates with both *time* and *tide*. Line 3 shows how semantically-related words are motivated by context, as here with *semantics* in the context of ‘a new dictionary volume’.

3.4 Semantically disambiguated information

Ambiguity is a central issue in automated text search. The fact is that, in addition to the obvious issues of polysemy and homography, most terms are multi-referential or multi-contextual (Renouf, 1993a) in use, and thus liable to generate low-precision results unless this is controlled, for example by restriction of the textual domain, or by the accompaniment of some contextual or analytical (e.g. grammatical) filter. The WebCorp word filter does the latter, by allowing the searcher to require the presence (or absence) of a disambiguating word on the same page as the search term. This is a simple but often effective means of improving precision, as shown in Figure 6, where the polysemous search term *sole* is limited to its piscatorial sense by the simple selection of *fish* as a required contextual item via the word filter.

1. I expected to get a nice juicy **sole**.
2. Andy was no more impressed with our syllabus of oeuf mayonnaise, **sole** véronique and sauce Espagnole than I was.
3. Quotas to cut fishing for **sole** in the English Channel and anchovies in the Gulf of Gascogne, in south-west France, are also of concern.
4. An extra 1C rise in temperature pushes haddock, cod, plaice and lemon **sole** 200 to 400 miles north, according to the WWF.
5. I recall the most splendid Dover **sole** at Scotts in Mayfair, assisted by a quite magnificent premier cru Chablis

Figure 6: extract of WebCorp output for search term *sole*; context *fish* specified

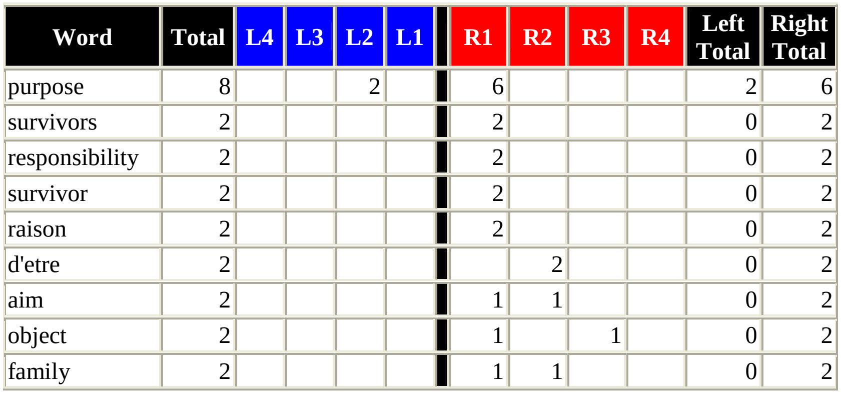
In Figure 7, the word *sole* is restricted to the sense of 'unique, only' by the word filter selection of the term *characteristic*. Curiously, the requirement for its presence somewhere in the text seems to licence the occurrence of some immediate collocates for *sole* which are compatible with but do not include *characteristic* itself – namely *purpose, aim, feature*. This indicates that the filtering word, if not functioning as an actual collocate, can function instead to create a semantic prosody (Louw, 1993) which encourages the desired sense of the search term to be realised. This fact of the language is convenient, if not entirely robust.

1. He and I once met for lunch for the **sole** purpose of continuing an argument
2. Katiek, what about a cause whose **sole** aim is to label people "evil" and "stupid"?
3. Its **sole** redeeming feature is that Stalin left their two-hour meeting complaining that Shaw was an awful person.
4. The **sole** black family on the vast Whinmoor estate in Leeds
5. Yesterday's summit finally dispelled the illusion that the UN is or can be the **sole** arbiter of war and peace.

Figure 7: extract of output for search term sole; context characteristic specified

3.5 External collocate profiles

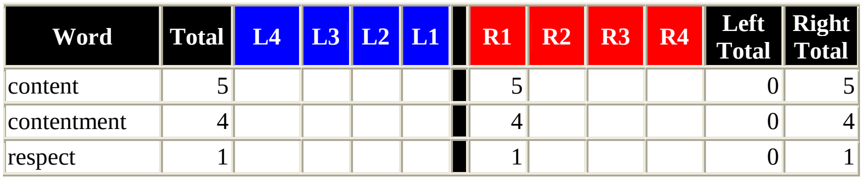
WebCorp also provides some basic statistical information, in particular about the 'collocational profile' (Renouf, e.g. 1993b) of the word. This is of necessity currently restricted to simple ranked frequency of occurrence in the set of pages visited. Figure 8, shows top-ranked 'external collocates' to complement the concordance lines in Figure 7, for the same search term, *sole*, again with the word *characteristic* in its presence. The slightly more extensive output shows that hypothesis that a single term can be used to focus context type certainly holds in this case: all top immediate collocates for *sole* here are compatible with the required sense.



Word	Total	L4	L3	L2	L1		R1	R2	R3	R4	Left Total	Right Total
purpose	8			2			6				2	6
survivors	2						2				0	2
responsibility	2						2				0	2
survivor	2						2				0	2
raison	2						2				0	2
d'etre	2							2			0	2
aim	2						1	1			0	2
object	2						1		1		0	2
family	2						1	1			0	2

Figure 8: external collocate output for search term *sole*; context *characteristic* specified

By way of further illustration, Figure 9 shows top-ranked ‘external collocates’ for the phrasal fragment [familiarity breeds], where the phrasal completive contempt has been suppressed by the word filter, and the word slot on which the query is focussed lies in position R1, outside the pattern submitted. Here, as shown in 3.3. above, phonology and semantics clearly play their role in the substitution.

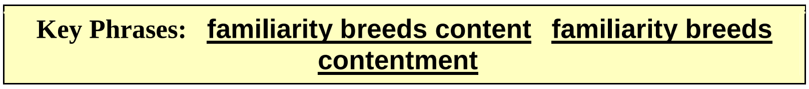


Word	Total	L4	L3	L2	L1		R1	R2	R3	R4	Left Total	Right Total
content	5						5				0	5
contentment	4						4				0	4
respect	1						1				0	1

Figure 9: top external collocates for search pattern [*familiarity breeds*], with phrasal component *contempt* filtered out

3.6 Key phrases

A simple heuristic (Morley, 2005) in WebCorp, involving a series of significant co-occurrence calculations, identifies a set of possible key phrases found within the results. In Figure 10, this reveals the more popular alternative phrases which emerge in place of the canonical when the key phrasal element *contempt* is suppressed.

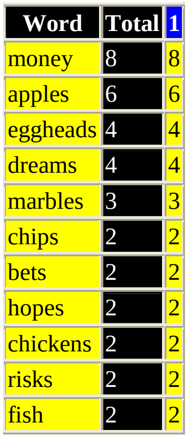


Key Phrases: **familiarity breeds content** **familiarity breeds contentment**

Figure 10: key phrases for search pattern [*familiarity breeds*], *contempt* suppressed

3.7 Internal Collocates

If a study is being conducted of lexical creativity within the phrasal pattern, WebCorp can provide the corresponding ‘internal collocate’ (Renouf, 2003) profile. This is illustrated in Figure 11 for the search pattern fragment *all your * in one basket*, where the internal collocates are non-hapax items which substitute for the suppressed *eggs* in wildcard position. These choices, as shown earlier, reveal some of the word play that characterises phrasal creativity in English.



Word	Total	1
money	8	8
apples	6	6
eggheads	4	4
dreams	4	4
marbles	3	3
chips	2	2
bets	2	2
hopes	2	2
chickens	2	2
risks	2	2
fish	2	2

Figure 11: top internal collocates within search pattern [*all your * in one basket*] with collocate *eggs* suppressed

3.8 Language detection

There are three main stages envisaged in the internationalisation of WebCorp (Renouf et al, 2004): handling/representing texts in other languages; refining search by specifying language; and automatic language identification. Of these, we have tackled the first two, since these have been prioritised by our users. The first is achieved by the integration of Unicode/double byte characters into the system. The second is accommodated through the selection of site domain (e.g. *.uk, .pt*), as a heuristic to control language or dialect variant, and it frequently works quite well, though it is not entirely reliable due to the well-documented cross-fertilisation which goes on between sites in terms of quotation of other languages, mirror-siting, and so on. Automatic language identification has been considered by us but not implemented as yet; it could be achieved by a combination of using the HTTP 1.1 language identification protocol, and by the implementation of one or other method of feature analysis. However, the true challenge comes not in identifying the language of a linguistically homogenous text, but of identifying words and short stretches in a different language within it. There is much knowledge already available in this area for us to draw on in the next stage of WebCorp.

4. Linguistic post-processing currently available with WebCorp

Post-processing of web-derived results adds time to what is already a slow procedure. Nevertheless, during 2002-3, we added post-processing options to WebCorp. One is the post-extraction alphabetical sorting of results on any specified collocate position. Another is the selection of desired and removal of unwanted concordance lines. We also added simple POS tagging, using the TNT tagger (internal version only).

An important move was the development of a means to conduct diachronic search. Web text protocols for dating are not applied consistently or at all, and at best they are ambiguous, so we devised a set of heuristics for searching for linguistic and other clues within the mark-up and the text itself, which have a measure of success in ordering results. Figure 13 demonstrates this for the word *radicalisation.*

18/08/1999 10:13:27 **1**	a widespread polarisation and	**radicalisation**	amongst the working class
16/09/1999 15:06:22 **1**	Kurds, Assyrians, Jews) and	**radicalisation**	of the Cossack movement
13/09/2001 18:19:39 **1**	Genoa – a new	**radicalisation**	has begun The 300
01/01/2002 00:00:00 **5**	has seen the increasing	**radicalisation**	of the Muslim position
24/01/2005 00:00:00 **2**	the areas of combating	**radicalisation**	and preventing terrorism.
09/07/2005 00:00:00 **3**	many factors behind the	**radicalisation**	of Muslim youth, including

Figure 13: post-extraction chronological listing of results for *radicalisation*

The first column here shows the date and time (where available) extracted by WebCorp for each of the originating web pages. This is followed by a number indicating the source of the date, where '1' is a server header date (the most reliable mechanism, '2' is a date metatag, '3' is a modification date in the body of the text, '4' is a copyright date and '5' is a date in the URL of the page (see Kehoe, 2005 for further explanation).

5. Remaining problems

As demonstrated, an extensive range of functions have successfully been developed for WebCorp, but given the intrinsic nature of web text, with its unorthodox definition of 'text', heterogeneity of data, lack of reliable punctuation and so on, several of these embody interim solutions and heuristics and could benefit from further improvement. Current WebCorp performance also lacks the high degree of processing and storage which is required to meet user needs expressed for simultaneous use for more users, including class-sized groups; grammatical and better collocational analysis; and more sophisticated pattern matching.

The primary constraint on the improvement of WebCorp performance, however, is its reliance on a commercial search engine. The problems posed by this dependence are as follows:

- the amount of web text searched is limited by time constraints, so that recall can be poor
- the proportion of potentially relevant web texts that is actually searched is limited (by search engine search criteria such as 'relevance' ranking and the 'indexability' (linking status) of a text), so that
- a similar small crop of texts is accessed each time, and a given search term garners largely the same results (although not reliably so, in terms of reproducibility), due to time-out and misjudged search prioritisation;
- the speed of results is inhibited

The delay built in by Google-dependent text extraction means that the time required for the linguistic post-processing of text is prohibitive, whether for POS tagging, for date and alphabetical sorting, or other requisite procedures. There are also unpredictable changes in Google service and even at the best of times, Google is geared to commercial rather than linguistic or even academic requirements. As discussed recently on 'Corpora-list' [http://torvald.aksis.uib.no/corpora/2005-1/0191.html], this can mean, for example,

- unreliable word count statistics
- limited and inconsistent support for wildcard search

With an eye to the long-term sustainability of the WebCorp system, we collaborated in 2001-2 with a UK-based search engine company, searchengine.com, who in exchange for linguistic information from us, provided first-hand experience of search engine technology and back-door access to their indexes, which speeded up response time.

6. The WebCorp Linguistic Search Engine

Our response to the problems anticipated and cited above has been to develop WebCorp with an eye to creating components that can be integrated into an independent, linguistically tailored search engine. We are currently calling this the 'WebCorp Linguistic Search Engine', since WebCorp functionality will be integrated into the new architecture alongside the search engine, and the whole fronted by an enhanced version of the WebCorp GUI. The new architecture is displayed graphically in Figure 7. The generic term 'linguistic search engine' is in fact a misnomer, since the search engine, while informed by linguistic knowledge, will not be 'linguistic' as such.

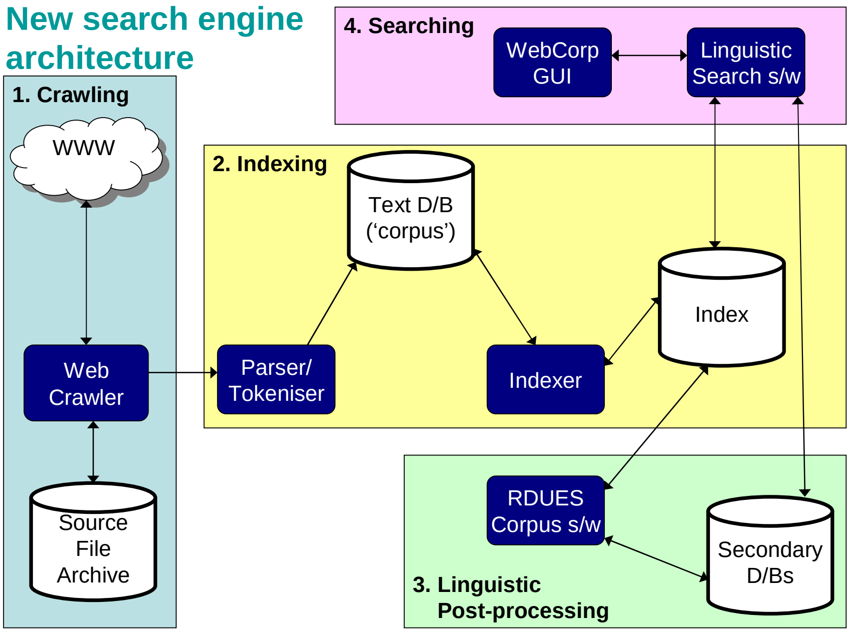


Figure 14: the new WebCorp Linguistic Search Engine architecture

The components of the new linguistic search engine system are as follows:

- web crawler
- parser / tokeniser
- indexer
- WebCorp tools
- WebCorp user front end
- more, also off-line, linguistic processing tools

and we have so far developed them individually, as we shall now outline.

6.1 Web crawler

Some five years ago, we developed a crawler module in Perl to select and download articles from UK newspaper websites. These are currently restricted to the Guardian and Independent, with whom we have had special arrangements. We shall now supplement them with tabloid and other categories of journalism. Not all newspaper sites have full archives like the Guardian, so instead of downloading them retrospectively, as we have done hitherto, we shall download the current day's articles daily, to build up the corpus progressively. Our initial estimate is that the newspaper 'domain' accessible through the WebCorp Linguistic Search Engine will contain at least 750 million word tokens. Our newspaper crawler incorporates the following features:

- exclusion lists (i.e. kinds of pages on newspaper sites NOT to download)
- error logging and re-queuing of failed pages
- extraction of date, author, headline and sub-headline
- URL parsing to extract section of newspaper (Sport, Media, etc)
- storage of articles by date (to facilitate diachronic analysis)
- removal of advertising banners and links ('boilerplate')
- stripping of HTML mark-up

We shall continue to use our tailored crawlers for our newspaper 'domain' and for other domains where pages are in a uniform format. We also have a specialised tool to extract neologisms from online articles in real-time. We shall expand this 'live' system to monitor and record neologisms, although once the web texts are downloaded into corpus format, we will begin to achieve this through the application of our APRIL system [http://rdues.uce.ac.uk/april.shtml], as we have begun to do with Guardian articles more recently.

In addition to structured sub-domains, we shall download a very large (multi-terabyte) subset of random texts from the web, to create a mini version of the web itself. Some users will prefer to look at this, much as they do with WebCorp at present than at particular sub-domains. The aim will not in itself be to build either specific sub-corpora or 'collections' of texts from the web, as other people such as Baroni & Bernardini (with BootCaT, 2004) have done, but to find a useful balance and combination of raw data, for instance in selecting random texts within a specific domain.

More generic tools will be required for the creation of this multi-terabyte mini-web, to cope with a variety of page layouts and formats. Several ready-made tools are available freely online but we are developing a new crawler for our specific task, building upon our experience with the newspaper downloads and making use of other open-source libraries whenever possible.

The new crawler will need to be provided with web addresses of where to embark on its crawl of the web (or 'seeded'). The search process could be completely random. This will not be appropriate, however, when building a structured corpus with carefully selected sub-domains. We shall employ other 'seeding' techniques including the use of Open Directory index, where editors classify web pages according to textual 'domain'. Our crawler will make use of the freely downloadable Open Directory 'RDF dumps' (http://rdf.dmoz.org/), containing lists of URLs classified by domain (or 'category'). We shall also consult human experts, university colleagues from our own and other disciplines, current WebCorp users and other contributors, to catalogue the major target websites in their field, so that these can also be used to seed the crawler. Thus a carefully planned seeding strategy will ensure a well-balanced and linguistically informed corpus.

New features of the crawler which are being developed include better duplicate detection: methods of comparing newly-encountered pages with those already stored in the repository to identify both updated and mirror versions of pages. We are also determined to improve on the date detection mechanism we have already created. Knowledge as to when our crawler first encountered a page may provide a clue as to when it was created, and the discovery of a new version of a page already stored will reveal when it was updated. The existence of our own independent search engine will allow us to conduct date detection off-line, not in real-time as at present. We shall also be able to classify changes and updates from the linguistic perspective, by scrutinising the page for changes in actual content rather than simply in mark-up or layout. Another area on which we have done considerable work, but which we should still like to improve on, is language detection, which could be done by the crawler or at the indexing stage.

6.2 Indexing

The source files will be stored in our standard RDUES format and then processed using specially adapted versions of the parsing, tokenising and indexing software which we have developed over the past 15 years, and run on texts downloaded from newspaper websites for the past 5 years. This will construct the corpus as a series of binary files and indexes. Our past experience indicates that we will be able to store 10 billion word tokens per terabyte of disk storage, including the processed corpus, indexes, raw HTML files (the 'source file archive' in Figure 14) and the secondary databases resulting from the linguistic post-processing stage outlined below.

Corpus updates will be incremental. New articles will be added to the newspaper domain daily, while other domains and the large mini-web 'chunk' will be updated at monthly intervals. Corpus processing will take place off-line and the new version of the corpus will 'go live' when processing is complete.

A constantly growing corpus could potentially cause problems when scholars attempt to reproduce previous experiments but find that the corpus composition has changed since in the meantime (cf. section 7.2 concerning frequency counts and statistics). For this reason, there will be a mechanism in the WebCorp Linguistic Search Engine allowing users to restrict searches to a specified subset (or 'collection') of texts which can be saved across sessions.

6.3 Linguistic post-processing

We shall be able to run on web texts any of the gamut of tools we can run on our current newspaper corpus. Where necessary, we shall also develop new tools, to provide a comprehensive range of corpus-processing functions. A priority is to exploit the tools created in major projects over the last 15 years, including those which generate collocates, 'nyms' (alternative search terms in the form of sense-related items), neologisms, summaries, document similarity measures, domain identification and so on. The sharing of these specialist language facilities will be a matter of individual negotiation: we shall be looking for relevant collaborative research proposals from potential users.

6.4 Searching

We shall develop new user interfaces, building upon our experience with WebCorp and other tools, such as the step-by-step and advanced APRIL neologisms demos [http://rdues.uce.ac.uk/aprdemo], taking into account user feedback, and so on.

All results from the linguistic post-processing will be stored in the secondary databases shown in the Figure 14 diagram of system architecture, and there will be new linguistic search software created to access the secondary databases.

7. Features and benefits of the new tailored web-search architecture

7.1 Increased speed

The system will now function as quickly as Google, but will be able to offer more functionality from a linguistic perspective. In terms of enhanced text quality, there will be a far greater rate of accuracy in respect of duplicate detection, sentence identification and full-text search. Text specification will be significantly improved with regard to domain detection, better date information for diachronic study and reliable language identification. Text search routines will be made more sophisticated with regard to specific domain search and specific URL sets.

7.2 Improved statistics

The web data will no longer be a vast, unquantifiable sea from which the system plucks an amount of data that cannot be evaluated in terms of its significance. The sub-web, or rather webs, which are regularly downloaded will be known entities, and thus reliable statistical counts and measures will be possible – in particular, the current WebCorp limitation to simple frequency counts will cease, and calculation of relative frequency and significance of phenomena such as collocation will commence.

7.3 Improved search

WebCorp search functionality will be vastly improved. This will include wildcard-initial search; wildcard matching for a variable number of intra-pattern search words up to a maximum span; POS specification, and lexico-grammatical specification.

8. Indicative output from the WebCorp Linguistic Search Engine

There follow some invented examples of the more complex and comprehensive linguistic and statistical analyses that we shall provide for the user once the WebCorp Linguistic Search Engine is up and running, and the post-processing operation will no longer be prohibitively time-consuming. The first two concern refined wildcard pattern search.

8.1 Wildcard-initial words as search terms

Google does not consistently support wildcard pattern search, and when it does, it does not allow wildcard-initial words as search terms. Our system will provide such information, as shown in invented output for the term [**athon*[in Figure 15. In addition, it will continue to be possible to specify textual domain (here 'UK broadsheets'), context length (here 'sentences') and dates (Sept-Dec, 2004).

1. including an ice-cream **scoopathon** and sausage treasure hunt
2. We left home at 10.15 to participate in the annual Right-to-Life **Walk-a-thon**
3. everyone talks about Zellweger 's **eatathon** as if she 'd been forced to lose a kidney
4. he strutted and tried to look feistier than he managed in that first **scowlathon**
5. A small army of people descend with cleaning materials for a five-hour **scrubathon**
6. It is one of the 20 tracks on the new **Spearsathon**
7. in October she will publish 'Manners', in time for the mass British **incivilityathon**

Figure 15: invented results for wildcard-initial search pattern [**athon*]

8.2 Variable number of words in wildcard position

For a search allowing for variation in number of words in the NP, we shall be able to provide a pattern search wildcard which allows for a specified maximum number of words. For instance, in a study of the 'It-cleft' construction [*it was* + PN *(3) + *that*], the *(3) would be a specification of all words in the wildcard position up to a maximum of 3. This would allow a search to yield results such as those shown in Figure 15. In addition, once we have established a sub-web processing system, we shall be able to provide lexico-grammatical search of the kind indicated by the search in Figure 16, where a combination of actual lexical realisations and grammatical categories may be specified.

(1 word)
1. it was **Blair** that gave him this power
2. it was **Iraq** that started the 1980-1988 war with Iran.

(2 words)
1. it was **Mitchell's reporting** that helped lead to the guilty verdicts
2. it was **Dennis's enthusiasm** that sparked the project to life.
3. It was **Seattle Weekly** that broke the Strippergate story
4. it was **the USA** that helped protect Australia from the Japanese

(3 words)
1. it was **the Muslim community** that could do things itself
2. it was **the Liberal Party** that ended the racist White Australia
3. it was **Cristijan Albers' Minardi** that punctured a left rear and went off the road

Figure 16: invented results for pattern [*it was *(3) that*], with a maximum of 3 words in wildcard position

9. Additional linguistic applications

9.1 Alternative search terms

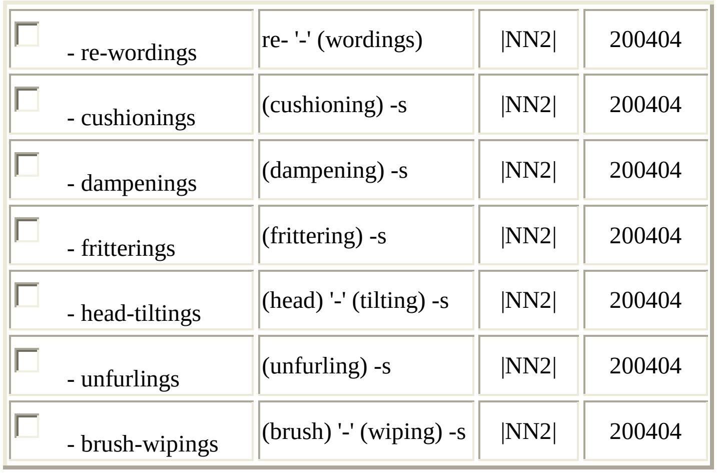
As said, the new search engine will allow us to bolt on past automated systems of linguistic analysis, of which we shall illustrate just two here. One such is the ACRONYM (Renouf, 1996) system, which automatically provides Wordnet-type sense-related synonyms or alternative search terms, and thus offers an opportunity for increase in recall. A sample of the ranked output it produces from our Independent/Guardian database is presented in Figure 17 for the terms *quest* and *questioned*, respectively:

- *quest*: pursuit, search, struggle, odyssey, ambition, endeavour, crusade, obsession, dream, mission
- *questioned*: quizzed, doubted, disagreed, examined, challenged, queried, argued, protested, speculated, lambasted

Figure 17: sense-related synonyms / alternative search terms from ACRONYM

9.2 Morphological Analysis

We also intend to append the APRIL (Renouf et al, forthcoming) project morphological analyser to the new system. This conducts the morphological analysis of target words, as well as providing plots of the fortunes of the target word or words across time. Figure 18 presents an extract of morphological information of the kind available with the new system – here, new nouns ending in suffix –*ings*, (or rather, nouns occurring for the first time since 1989). Figure 19 presents a time plot to allow the examination of the productivity patterns of the suffix –*esque*, which reveals a slow growth, peaking in 2004.



□ - re-wordings	re- '-' (wordings)	|NN2|	200404
□ - cushionings	(cushioning) -s	|NN2|	200404
□ - dampenings	(dampening) -s	|NN2|	200404
□ - fritterings	(frittering) -s	|NN2|	200404
□ - head-tiltings	(head) '-' (tilting) -s	|NN2|	200404
□ - unfurlings	(unfurling) -s	|NN2|	200404
□ - brush-wipings	(brush) '-' (wiping) -s	|NN2|	200404

Figure 18: extract of APRIL results from April 2004: new nouns with suffix -*ings*

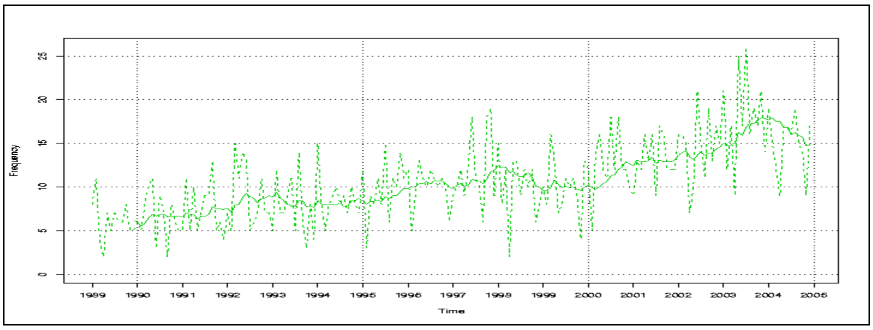


Figure 19: APRIL time plot showing number of new words with the suffix -*esque*

10. Concluding remarks

There is general agreement among those who have devised and implemented automated methods of extracting linguistic research data from the web via a commercial search engine (e.g. us; Fletcher, 2001 & this volume; Resnik & Elkiss, 2003), as well as among reviewers of such initiatives (e.g. Lüdeling, Evert and Baroni, this volume; Kilgariff, 2003), that the mediated route is less than ideal, particularly for on-line retrieval systems. Having foreseen this problem at the outset, we have worked steadily over the last five years to develop the components required to create a linguistically-tailored and accessorised search engine, and we shall in the coming months assemble an infrastructure that will be progressively incorporated into the WebCorp front-end. We are confident that this will enhance its performance on the fronts outlined above, and allow it to support serious research. Improvements to our system will be incrementally perceptible at http://www.webcorp.org.uk/.

11. References

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

Fletcher, W. (2001) 'Concordancing the Web with KWiCFinder', in *Proceedings of The American Association for Applied Corpus Linguistics Third North American Symposium on Corpus Linguistics and Language Teaching*. Available online from http://www.kwicfinder.com.

Kehoe, A. (2005) 'Diachronic linguistic analysis on the web with WebCorp', in A. Renouf and A. Kehoe (eds.) *The Changing Face of Corpus Linguistics*. Amsterdam & Atlanta: Rodopi.

Kehoe, A. and Renouf, A. (2002) 'WebCorp: Applying the Web to Linguistics and Linguistics to the Web'. In *Online Proceedings of World Wide Web 2002 Conference, Honolulu, Hawaii, 7-11 May 2002*. http://www2002.org/CDROM/poster/67/

Kilgarriff, A. (2003) 'Linguistic Search Engine'. *In Proceedings of The Shallow Processing of Large Corpora Workshop (SProLaC 2003) Corpus Linguistics 2003,* Lancaster University.

Louw, B. (1993) Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In Baker, M., Francis, G. & E. Tognini-Bonelli (eds) *Text and Technology*. In Honour of John Sinclair. Philadelphia/Amsterdam: John Benjamins.

Morley, B. (2005) 'WebCorp: A tool for online linguistic information retrieval and analysis', in Renouf, A. and A. Kehoe (eds.) *The Changing Face of Corpus Linguistics* (Amsterdam & Atlanta: Rodopi).

Renouf (1993a) 1993: 'What the Linguist has to say to the Information Scientist', in Gibb, Forbes (ed.): *The Journal of Document and Text Management*, vol. 1:2, 1993. 173-190

Renouf, A. (1993b) ‘Making Sense of Text: Automated Approaches to Meaning Extraction’, in *Proceedings of 17th International Online Information Meeting, 7-9 Dec 1993*. 77-86.

Renouf, A. (1996) ‘The ACRONYM Project: Discovering the Textual Thesaurus’, in I. Lancashire, C. Meyer and C. Percy (eds.) *Papers from English Language Research on Computerized Corpora (ICAME 16)* Amsterdam & Atlanta: Rodopi. 171-187.

Renouf, A. (2002) ‘WebCorp: providing a renewable data source for corpus linguists’, in Granger, S. and S. Petch-Tyson (eds.) *Extending the scope of corpus-based research*. Amsterdam & Atlanta: Rodopi. 39-58.

Renouf, A., Morley, B. and A. Kehoe (2003) ‘Linguistic Research with the XML/RDF aware WebCorp Tool’. In Online Proceedings of *WWW2003*, Budapest. http://www2003.org/cdrom/papers/poster/p005/p5-morley.html

Renouf, A., Kehoe, A. and D. Mezquiriz (2004): ‘The Accidental Corpus: issues involved in extracting linguistic information from the Web’, in Aijmer, K. & B. Altenberg (eds.) *Proceedings of 21st ICAME Conference, University of Gothenburg, May 22-26 2002*, Amsterdam/Atlanta GA: Rodopi. 404-419

Renouf, A., Pacey, M., Kehoe, A. and P. Davies (forthcoming), ‘Monitoring Lexical Innovation in Journalistic Text Across Time’

Resnik, P. and A. Elkiss (2003). ‘The Linguist's Search Engine: Getting Started Guide’. Technical Report: *LAMP-TR-108/CS-TR-4541/UMIACS-TR-2003-109, University of Maryland, College Park, November 2003*.