Diachronic linguistic analysis on the web with WebCorp

Andrew Kehoe

University of Central England in Birmingham

Abstract

The WebCorp project has demonstrated how the Web may be used as a source of linguistic data. One feature of standard corpus analysis tools hitherto missing in WebCorp is the ability to filter and sort results by date. This paper discusses the dating mechanisms available on the Web and the date query facilities offered by standard Web search engines. The new date heuristics built into WebCorp are then discussed and illustrated with a case study.

1. Introduction

'For modern corpus linguists, diachrony is typically the study of change in one or more aspects of language use just within (or across) a timespan of 10-30 years' (Renouf, 2002). There are, however, some language changes which are too recent to be evidenced in standard corpora and the WebCorp project (http://www.webcorp.org.uk/) was set up to treat the Web as a corpus from which such linguistic information can be extracted. (See Renouf, Kehoe and Mezquiriz, 2004, for further background on WebCorp.)

The Web is also useful as a linguistic resource when searching for words or phrases too rare to appear in any standard corpora. Bergh, Seppänen & Trotta (1998) were among the first researchers to turn to the Web as a linguistic resource, searching for rare fronted-*which* constructions ('*x which are believed can y*', etc) using the AltaVista search engine. Our WebCorp usage logs show that new and rare constructions continue to be among the most common search terms entered.

When searching for new or rare constructions on the Web, it is essential to know the dates on which the Web pages from which examples have been extracted were written. This paper examines the dating mechanisms available on the Web, assessing their usefulness for linguistic analysis and describing how the WebCorp system has been adapted to support diachronic analysis.

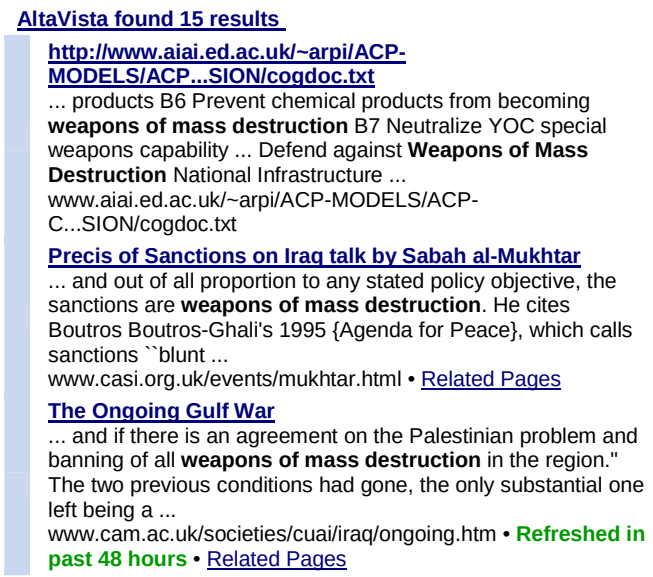
2. Searching by date on the Web

Standard Web search engines are surprisingly limited when it comes to date-restricted queries and, indeed, the Web itself lacks the necessary means for recording either temporal and diachronic information. We ran tests to discover what dating mechanisms are available on the Web and found that the only

Final version to appear in A. Renouf & A. Kehoe (eds.) *The Changing Face of Corpus Linguistics,* Amsterdam: Rodopi (2006), pp. 297-307.

potentially reliable mechanism is the 'Last Modified' header which is passed to the client when a page is requested from a Web server. This records the date on which the page was last saved by its author, although our tests have shown that only just over half of the pages returned by the Google search engine include this header when accessed directly (Kehoe & Renouf, 2002). Often, dynamically-generated pages do not return this header, and some Web servers are configured not to return it at all.

Some Web search engines do offer date-restricted queries. Google allows queries to be restricted to the past 3, 6 or 12 months, but this is not sufficient for linguistic research. Taking the phrase '*weapons of mass destruction*', which became widely used in early 2003, a linguist may wish to search for the earliest occurrence of the term on the Web. This is not possible in Google because the maximum date restriction is 'within the past 12 months' and the user cannot restrict the query to pages written *before* a certain date or *between* two points in time. AltaVista does offer a date span option on its Advanced Search page, and the 'weapons of mass destruction' query returns only 15 results when restricted to Web pages written between 1/1/96 and 31/12/97 (as opposed to 26,050 with no date restriction). However, the AltaVista results list does not show the authorship date of each page and, in most cases, it is not possible to find this date, even by clicking on the link and accessing the page itself.



AltaVista found 15 results

http://www.aiai.ed.ac.uk/~arpi/ACP-MODELS/ACP...SION/cogdoc.txt
... products B6 Prevent chemical products from becoming **weapons of mass destruction** B7 Neutralize YOC special weapons capability ... Defend against **Weapons of Mass Destruction** National Infrastructure ...
www.aiai.ed.ac.uk/~arpi/ACP-MODELS/ACP-C...SION/cogdoc.txt

Precis of Sanctions on Iraq talk by Sabah al-Mukhtar
... and out of all proportion to any stated policy objective, the sanctions are **weapons of mass destruction**. He cites Boutros Boutros-Ghali's 1995 {Agenda for Peace}, which calls sanctions ``blunt ...
www.casi.org.uk/events/mukhtar.html • Related Pages

The Ongoing Gulf War
... and if there is an agreement on the Palestinian problem and banning of all **weapons of mass destruction** in the region." The two previous conditions had gone, the only substantial one left being a ...
www.cam.ac.uk/societies/cuai/iraq/ongoing.htm • **Refreshed in past 48 hours** • Related Pages

Figure 1: Extract from AltaVista results for date-restricted *weapons of mass destruction* query (run on 24/06/03)

The last result in this extract (like two others of the 15) is marked as being "Refreshed in past 48 hours" but it is unclear why AltaVista would need to

update its record of a page which has supposedly not been modified for at least 5 years. It is clear that date queries in search engines do not always produce accurate results. When run in Google on 24/06/03 and restricted to 'the past 3 months', the '*weapons of mass destruction'* query returned over 1.3 million hits. However, when accessing each of the hit URLs directly, we found '*Last Modified*' headers containing the dates 20/03/02, 08/11/02, 11/07/01 and 27/07/00 amongst others, making it impossible that these pages were written, or even altered, in the past 3 months. Price & Tyburski (2002) have noted similar problems with date queries in search engines and suggest that there may be a bias towards the date on which a Web page was last indexed by the search engine, rather than towards the date it was written or last modified. This would be entirely unhelpful information for most purposes, but particularly so for linguists. Furthermore, we are not aware of any mainstream Web search engine which allows the **sorting** of results by date, a standard feature in corpus analysis software.

3. The implementation of diachronic queries in the WebCorp tool

We have adopted a multi-layered approach, using a range of sources to allow more accurate date-restricted linguistic analyses on the Web. The first step involves the examination of the server headers of a page, to discover whether the '**Last Modified**' date is present. This will be of the form:

```
Date: Tue, 15 Jul 2003 10:43:54 GMT
Accept-Ranges: bytes
ETag: "366136-c16b-3e0baf1f"
Server: Apache/1.3.12 Cobalt (Unix)
Content-Length: 49515
Content-Type: text/html; charset=iso-8859-1
Last-Modified: Fri, 27 Dec 2002 01:38:39 GMT
Client-Date: Thu, 07 Aug 2003 13:34:58 GMT
Client-Response-Num: 1
Proxy-Connection: close
Title: Texting
```

Figure 2: Sample 'Last Modified' header

If there is no 'Last Modified' header, the second step is for WebCorp to examine the **user-specified meta-tags** for a date tag of some sort. Through this method we discovered that, although the pages on the BBC News website (http://news.bbc.co.uk) do not return a 'Last Modified' header, they do include an 'OriginalPublicationDate' meta-tag. Pages on other sites contain similar metatags in various formats.

If neither a 'Last Modified' header nor a date meta-tag is present, a third heuristic is applied: WebCorp looks for a **user-specified modification date** within the body of the Web page. Such dates are even more variable than date

meta-tags, in terms of the date format and exact wording used: '*Last modified*:', '*Last update'*, '*last revis*ed', etc. A regular expression has been designed to match any date, no matter how it has been worded by the page author, and all dates are converted into a standard format.

The fourth heuristic, applied when a Web page passes through each of the three preceding stages with no date found, is to look for a **copyright date** on the page. WebCorp will match any form of the copyright symbol *(©, (c), copyright, copywrite*, etc), and where a range of dates is specified (e.g. '*(c) 2000-2001'*), the later date is taken. With copyright dates, the month and day are unknown, so WebCorp defaults to '*1 January'* of that year. Our tests have shown that a large proportion of the Web pages returned by search engine queries do contain a copyright date (between 50% and 70%, depending upon the search term used).

Our final heuristic is to examine the **URL** of a Web page for clues about the date on which it was published. Some sites, particularly news sites, archive pages by date, e.g. http://www.cnn.com/2003/WORLD/americas/01/05/venezuela shootings /index.html (which was published on January 5, 2003).

3.1 WebCorp user interface options for date search

The WebCorp user interface allows users to specify date restrictions in two different ways, either by selecting an option from a drop-down menu or by entering a date range. The drop-down menu allows the user to include only pages which are dated '*in the past month'*, '*in the past 3 months'*, '*in the past 6 months'*, '*in the past year'*, '*more than 1 year ago'*, '*more than 2 years ago'* or '*more than 5 years ago*', thus providing more precision than the Google date options. Alternatively, the user can choose to enter a date range and restrict the query to pages dated within that time period.

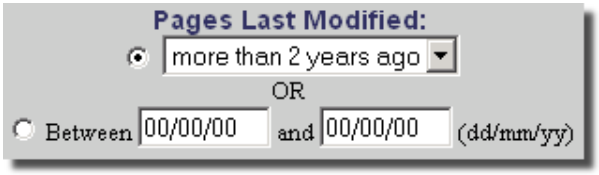


Figure 3: WebCorp date options

3.2 Sorting retrieved texts in date order

The WebCorp date module returns a date in a standard format (yyyy mm dd hh:mm:ss) and gives an indication of the type of date found (1:Server Header, 2:Date Metatag, 3:Author-Specified Modification Date in Document Body, 4:Copyright Date, or 5:Date in URL). The type of date is included so that the user can gauge how reliable a particular date is likely to be, and to allow secondary sorting on date type.

http://www.nwc.navy.mil/press/Review/1998/summer/bkr2su98.htm
Document Dated: 2003/06/25 15:29:18 (server header)
Plain Text Word List 683 tokens, 385 types
Ideally, **shock and awe** would both paralyze and deter an opponent before the bullets fly.

http://www.wsws.org/articles/2003/jan2003/war-j30.shtml
Document Dated: 2003/01/30 00:00:00 (metatag)
Plain Text Word List 1799 tokens, 891 types
US plans "**shock and awe**" blitzkrieg in Iraq

http://www.spiritualityhealth.com/newsh/items/newsitem/item_5541.html
Document Dated: 2003/01/01 00:00:00 (copyright)
Plain Text Word List 785 tokens, 431 types
Briefly, "**shock and awe**" refers to a military strategy that could be used in the threatened U.S. war on Iraq.

Figure 4: Enhanced WebCorp date output

WebCorp then allows sorting of concordance lines by date, in ascending or descending order, as shown in Figure 5. In the sorted output, a string is appended to the beginning of each line, showing the date, and the date type as a number from 1 to 5. The originating Web pages can be accessed by clicking on the keyword in bold red type.

16/04/2003 00:00:00 3	says he invented the term "	**shock and awe**	" but that the concept draws
25/06/2003 15:29:18 1	in blitzkrieg, rapid dominance produces	**shock and awe**	through four elements, including "rapidity
01/07/2003 00:00:00 2	months. It is time to	**shock and awe**	those potential customers--not with discounted
16/07/2003 10:29:00 1	its war plan—“	**shock and awe**	.” The notion is that
16/07/2003 10:29:00 1	his assessment that a “	**shock and awe**	” bombing campaign would crumble

Figure 5: Sorted WebCorp date output

3.3 Assessing the WebCorp date-identification heuristics

Figure 6 summarises the success rate of our date-identification heuristics for 8 different search terms. For each search term, we took the first 100 URLs returned

by Google and ran our heuristics on them. Previously, WebCorp had used only server header dates. The light shaded area shows the number of dates added by our new heuristics, culminating in the right-hand column with the total number of new examples of date information found. The 'Errors' category includes URLs which were returned by Google but could not be accessed by our tool at the time of the experiment, either because they no longer existed or because the server they were held on was temporarily inaccessible.

Word	No date	Error	Server Header	Metatag	Author-specified	Copyright	URL	Dates Added
texting	34	15	21	21	1	7	1	30
news	46	1	32	5	5	10	1	21
normalcy	31	7	37	7	1	8	7	23
phat	30	6	55	1	0	5	1	7
humongous	20	2	38	2	2	35	1	40
Liverpool	15	1	78	0	3	3	0	6
blogger	41	2	28	1	7	15	6	29
WMDs	15	8	39	0	1	27	10	38
Averages	29	5.25	41	4.63	2.50	13.75	3.38	24.25

Figure 6: Summary of dates added by WebCorp heuristics

As Figure 6 illustrates, our heuristics allow us to add date information for an average of 24.25% of Web pages. As well as increasing recall in this way, we are increasing precision by basing our date identification on known factors, in order of likely accuracy, rather than on unreliable search engine date options. There are, however, several issues regarding precision which must still be addressed.

4. Limitations of the new date heuristics & diachronic analysis on the web

4.1 Server header and metatags

The '*Last Modified*' date of a Web page will only correspond to the authorship/publication date if the file has never been re-saved. The problem with the Web is that there is no archiving mechanism or concept of 'editions' and, in most cases, when a text is modified the original version is lost forever. (There are some cases were versions of documents are carefully archived on the Web, and there are sites such as http://www.archive.org/ which attempt to keep a record of how individual Web sites looked at particular points in time, but these are not widespread or easily searchable for linguistic data.)

It may be the case that a Web page was written in 2001 but the author made a small alteration (perhaps correcting a typographical error) two years later, altering the 'Last Modified' date automatically in doing so. The altering of the copyright date on a Web page each year will also change the 'Last Modified'

header (see below). There are parallels here with plagiarism detection (i.e. small changes being made to existing documents at a later date) and perhaps work in this field could inform our work on Web date analysis.

As an extreme example of the difference between 'Last Modified' headers and the actual authorship dates of Web texts, the URL http://the-tech.mit.edu/Shakespeare/cleopatra/full.html returns the 'Last Modified' header 'Wed, 18 Oct 2000 20:58:44 GMT' yet the text on the page is Shakespeare's *Antony & Cleopatra*, written in 1606-7. This is something the user must be aware of when viewing dated concordance lines but it is, in a sense, equivalent to an edition of Shakespeare being published in 2000 as a 'new' book.

Date meta-tags provide more flexibility and allow page authors to include original authorship and publication dates in addition to last modification dates, as illustrated by the '*OriginalPublicationDate*' tags on the BBC News website. However, there are as yet no widely used meta-data standards and this vacuum encourages variation, with different sites using different tags. This makes it impossible for WebCorp to interpret all date meta-tags.

4.2 Author-specified revision date

Like '*Last Modified*' headers, author-specified revision dates in the body of a Web page indicate when the page was last changed but, for the most part, authors do not give details about exactly what was changed on the page on that date. Unlike '*Last Modifie*d' headers, these revision dates are not updated automatically when the page is altered, and it is left to the author to update them manually.

4.3 Copyright date

As discussed above, we found copyright dates on between 50 and 70% of Web pages. The problem is that the copyright date at the bottom of an individual web page may be a site-wide copyright date and not reflect the actual authorship date of individual pages on a site. Also, the copyright date on a page may be altered routinely each year, no matter whether the page has been otherwise modified. In some cases, page authors post-date copyrights – the page at http://sozluk.sourtimes.org/show.asp?t=terminatrix, for example, has a copyright date of 1999-2012.

However, we use copyright dates only as a fallback heuristic measure and, in cases where other methods fail, they can provide a useful estimate of the date of a Web page. Copyright information can usually be relied upon to provide a point in time *after* which a page must have been authored if nothing else.

4.4 Date in URL

This dating heuristic can be useful although, again, formats vary between sites and we place this heuristic last in line, as we feel our other detection techniques are more reliable. For example, we have encountered some Web pages with

'*1945*' in the URL, where this refers to the year which is being discussed rather than to the year in which the page was written. Even limiting the date window to 1990-2100 produces some errors, caused by 4 digit numbers in URLs which are not years.

5. Diachronic analysis?

Renouf's definition of diachrony involves a time-span of only 10-30 years – really 'brachychrony' (Renouf, 2002). On the Web, there are very few pages that are more than 10 years old, as it was only in 1994/5 that the Web began to grow in popularity. There is also a bias towards new texts in Web search engines. Google does have a searchable index of newsgroup posts dating back to 1981 but, while these posts are useful and cover a wide variety of topics, the genre is limited to 'discussion group' and does not offer the same range of texts as the Web.

The Web is, however, a valuable resource to supplement the analysis of linguistic change *within* a 10-year period, as the following case study illustrates.

6. Case Study

The aim in this section is to trace the introduction of the word *alcopops* into the English language, a word referring to drinks, marketed at young people, which are a blend of alcohol and 'pop', or carbonated, fruit flavoured liquid. This word was coined in the mid- to late-1990s but does not appear at all in the BNC World Edition, either in singular or plural form.

Google returns over 17,000 hits for the term *alcopops* but, as discussed above, it is not possible to restrict Google queries by date, other than to occurrences in the past 3, 6 or 12 months. AltaVista returns only 41 results when the *alcopops* query is restricted to the time-span 01/01/94-31/12/99 but, again, there is no way for the user to sort the results by date or see the date assigned to any of these 41 pages and even clicking on the link will not show the date if it is in the server header or metatags.

In contrast, using WebCorp (with Google selected as the Search Engine option) the linguist is able to extract 472 concordances from 200 Web pages for the term *alcopops* and view these in ascending date order. The full date-sorted results (as run on 06/08/03) can be found at http://rdues.bcu.ac.uk/alcopops.html. The extract in Figure 7 illustrates the earliest occurrences of the term (after the pages with unknown dates)

The first example in Figure 7 appears to be from 1995, but the 5 at the end of the date string indicates that this date was extracted from the URL of the page: http://www.bbc.co.uk/cult/ilove/years/1995/fashion1.shtml. This page discusses the news and fashions of 1995 but derives from the 'nostalgia' section of the BBC website and was not actually written in that year. Since the page contains no Last Modified header, no date metatags, no user-specified modification date and

no copyright date, there is no way of discovering its actual authorship date. However, this concordance does tell the linguist that the term *alcopops* was first introduced into the UK in 1995.

There are then several contexts from 1997, from two websites where the dates displayed are definitely accurate. The first six of the 1997 concordance lines are from http://www.allaboutbeer.com/news/world/97alcopop.html (UK news from a US-based brewing industry site) where we see that *alcopops* is introduced in double quotes and defined as "*the popular fruit-flavored alcoholic drinks*", an orthographic convention indicating that, although the drinks themselves are 'popular' in the UK by this stage, *alcopops* is still seen by the author as a new term. It is also new to his American readers, as the drinks have recently been "*rolled-out"* in the United States (5th concordance line from that site).

The remaining two 1997 concordances are from a site in New Zealand (http://www.nzdf.org.nz/update/messages/33.htm) and again the word *alcopops* is presented in quotes and defined by the author, this time as "*pre-mixed alcoholic drinks"*.

The last context in this extract provides an example of the word *alcopops* being used in 1998 on a page written in Spanish (http://www.msc.es/salud/epidemiologia/resp/199801/editorial.htm), with the date extracted from the server header, the most reliable mechanism. By 2003 (see the full output on the RDUES website), there are examples of *alcopops* appearing on native-language sites in France, Denmark, Belgium and Switzerland, among others, indicating that the word has been borrowed by other European languages.

01/01/1995 00:00:005	A huge furore erupted when lemonade-flavoured **alcopops** like 'Two Dogs' and 'Hooch' hit the shelves in 1995.
01/01/1997 00:00:004	Although British companies that sell " **alcopops** " changed their advertising after complaints they target underage drinkers, the popular fruit-flavored alcoholic drinks remain under fire.
01/01/1997 00:00:004	A study discovered that 84 percent of boys and 87 percent of girls 15-16 years old who drink alcohol at least once a month **alcopops** have tried .
01/01/1997 00:00:004	More than 6 in 10 wrongly believed that **alcopops** were less strong than beer.
01/01/1997 00:00:004	More than half of those 15-16 years old who were surveyed said **alcopops** were "something to drink with our friends at parties."
01/01/1997 00:00:004	The booming success of British brands such as Hooper's Hooch and Two Dogs has led to a rollout of **alcopops** in the United States.
01/01/1997 00:00:004	Those 13-16 said **alcopops** were the easiest alcoholic drink to obtain.
27/06/1997 00:30:031	Tighter rules to control ' **alcopops** '
27/06/1997 00:30:031	Fifteen liquor companies have voluntarily agreed to tighten controls on the naming, packaging and marketing of " **alcopops** ", pre-mixed alcoholic drinks, after controversy over their appeal to under-age drinkers.
24/02/1998 16:19:471	having regard to the widespread appeal of ' **alcopops** ' and 'designer drinks' to children and young people which has been confirmed by recent independent research such as published in the British Medical Journal, B
24/02/1998 16:19:471	Calls on the Commission to introduce European-wide guidelines for the promotion, marketing and retailing of **alcopops** and designer drinks;
17/03/1998 08:59:171	Las bebidas **alcopops** , con especial referencia a los refrescos con alcohol, se han convertido en un gran éxito de ventas, principalmente entre los jóvenes en el Reino Unido

Figure 7: Date-sorted WebCorp output for the term *alcopops*

7. Further work

As the case study has illustrated, our heuristics allow diachronic linguistic analysis on the Web in a way which is not possible when using standard Web search engines. There are, however, some enhancements which could be made. The first would be to use the hyperlink structure of the Web to aid the dating of individual pages. At a simple level, if a definite authorship date for a page (Page A) is known, and Page A links to another page (Page B), this places the original authorship date of Page B at some point in time *before* that of Page A. Similarly, if Page B links to a third page (Page C), this places the authorship date of Page B at some point in time *after* that of Page C. Complex networks of dating information could be built using this method.

It would also be possible to conduct feature analysis on Web pages to estimate authorship dates. One level of analysis would be to look for the latest dates in the bibliography sections of online books and academic papers. A more complex task would be to look for key names and events mentioned, as clues to authorship date. The names of presidents and prime ministers, or references to events such as *9/11*, etc could be used to establish authorship dates as being after a certain point in time. Work in the field of forensic linguistics may be helpful here.

The hope is that Web dating mechanisms will improve in the future to allow more accurate dating of pages. The Resource Description Framework (RDF - http://www.w3.org/RDF/), put forward by the World Wide Web Consortium as a metadata standard, may go some way towards achieving this, by allowing the page author to include a qualifier to specify exactly what the 'date' included in an XML document header represents: whether 'Created', 'Valid', 'Available', 'Issued' or 'Modified' (Kehoe & Renouf, 2002). This goes beyond the somewhat limited 'Last Modified' header system that is in place at present.

Acknowledgements

The WebCorp project was funded by the EPSRC, and would not have been possible without the software development expertise of Jay Banerjee, David Mezquiriz, Barry Morley & Mike Pacey. I am grateful to the anonymous reviewer for pointing out the parallels between our work on Web date detection and work on plagiarism detection and forensic linguistics.

References

Bergh, G., A. Seppänen and J. Trotta (1998), 'Language Corpora and the Internet: A joint linguistic resource', in: A. Renouf (ed.) *Explorations in Corpus Linguistics*, Amsterdam/Atlanta, GA: Rodopi

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

Price, G. and G. Tyburski (2002), 'It's Tough to Get a Good Date with a Search Engine', in: *SearchDay*, June 5 2002, http://www.searchenginewatch.com/searchday/article.php/2160061

Renouf, A. (2002), The Time Dimension in Modern English Corpus Linguistics', in: B. Kettemann and G. Marko (eds.). *Teaching and Learning by Doing Corpus Analysis. (Papers from the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July 2000)*, Amsterdam/Atlanta, GA: Rodopi

Renouf, A., A. Kehoe and D. Mezquiriz (2004), 'The Accidental Corpus: Some Issues in Extracting Linguistic Information from the Web', in K. Aijmer and B. Altenberg (eds.) *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, Amsterdam/Atlanta GA: Rodopi