Using the World Wide Web as a Massive Corpus

Martin Edwardes
University of East London
martin.edwardes@btopenworld.com

Paper presented to the British Association for Applied Linguistics conference
Cork, September 2006

Contents

Introduction

The World Wide Web is a recent creation. A few weeks ago it celebrated its fifteenth birthday, and it's already in its third generation. Tim Berners-Lee, the creator of the Web, probably did not anticipate the direction and speed of social change that the Web has both represented and encouraged.

In its first incarnation the Web became a large database of human carnality. This is likely to be because the early adopters were what we would call nerds and geeks, for whom sex with someone else in the room was an occasion. The sexuality of the early Web was largely an innocent, one-handed thing, and it was only with the later adopters that the more sinister aspects of pornography took over.

By then we were into the second incarnation, and the Web had become a marketplace. Software allowing credit card transactions turned the Web into the biggest department store ever – anything and everything was available. However, rumours of the death of the High Street proved to be somewhat exaggerated.

In its third incarnation, the Web is finally becoming what Tim Berners-Lee envisioned it to be: a free-form information warehouse with a worldwide range of contributors. The popularity of the blog has given non-experts a presence on the Web. All that is needed now for a web presence is a willingness to write or design; the technical details can all be handled by widely available – and largely free – software.

On the Web, new uses have not replaced the old. Instead, the number of pages has increased steadily over the years, covering a range of new uses and subjects. The pornography is still there, but it is a tiny part of what is now available. Currently there are estimated to be more web pages than there are people on Earth. About half of them are in English, which gives a corpus of close to a trillion English words, or the equivalent of ten million books. This has to be considered a massive corpus.

What the Web is

What exactly is the Web? Simply put, it is a massive, open-ended but hierarchical database. Each record consists of a web page, and what goes on that web page is theoretically unlimited. If it can be reduced to a string of binary digits, the Web can handle it. Groups of web pages make a site, which usually has an identifiable owner responsible for site content; and all the sites together make up the Web.

However, the Web is more than just a repository, it is a tool of communication. Pages are set up in the expectation that they will be read, so almost every page is the opening utterance in a dialogue. This dialogue might be continued via email, directly on the web page, or via other web pages produced by other people on other sites. Every time you encounter an offsite link, it is an utterance in the massive dialogue that the World Wide Web has become.

Of course, any corpus is only as effective as the search engine which accesses it; and, fortunately, with the World Wide Web we have a series of fast, efficient and free search engines. They are not specifically designed for linguistics use, and have some limitations. Some words are not indexed at all – for instance, the frequency of *the* is impossible to determine, no search engine will give you a count. It is also pointless trying to work out absolute frequencies for words or phrases. The absolute number of words on the Web is unknown, so there is no total word count from which ratios can be calculated.

This leaves relative frequencies of word and phrases, and the Web is excellent for producing datasets for this type of analysis. To give a trite example, AltaVista produces 21 million hits for *linguistics*, 27 million for *anthropology*, and 35 million for *sociology*. These are all dwarfed by the 101 million hits for *psychology*. As a subject, popularity is obviously not our strong suit.

Using the Web

So what are the tips to getting the most out of the Internet? The search engine to use is not a problem, they are all good enough for our purposes, and they all work in much

the same way. In the early days of the web, search engines referenced on the text in page headers, in page tags or in the first few lines on the page. Nowadays, the biggest engines - Google, Yahoo, MSN and AOL, which together handle about 90% of searches – all index the content of pages as well. Additionally, they index the content of documents in Word, pdf and several other formats. You can find all these search engines by typing their names into the large box at the top of the Explorer screen.

To search for a particular word you just type it into the FIND box, and click on the FIND or SEARCH button. If, however, you wish to search for a phrase then you should type it within quotes. **Appendix 1 – Making Searches** sets out in more detail the rules for using search engines.

When you get your results you will see somewhere on the page the number of web pages found. This does not count multiple occurrences on the page, only pages found. For instance, a search on *"I will be finding"* on AltaVista on 18th August 2006 found 14,300 occurrences. To compare, Google found 19,500, AOL found 1,590, Yahoo found 14,600 and Excite found 64. When I did this search on AltaVista in 2003 I found only 197, which shows how much the Web has expanded. This illustrates the importance of using only one search engine for a project, and the need to extract statistics within short time frames. Web searches are not comparable over time, nor between different search engines.

You've now got a rough frequency count for the word or phrase you are interested in; what do you do next? You should check a few sites (the first 40 or 50 should provide a statistically accurate estimate). Click on the site name to view it, and use BACK in the browser to return to your list of prospective sites. Use EDIT/FIND (ON THIS PAGE) to find the occurrences of the selected text on the page. You will find sites have been incorrectly included in the search for several reasons:

- The page may have been removed or be temporarily offline. Unfortunately you can't tell which it is from the message. This is often referred to as a 404 page, because that is the official Web error number for a link that takes you nowhere.
- The site text may have been changed, and the search engine compressed database has not been updated. This often happens with news sites, where the page content is changed on a regular basis and the old content is archived.
- The site text is not exactly as you selected. The selection will ignore a lot of punctuation, so a search for *hope for* will find *…at least, that is what I hope. For tomorrow, I am….* However the FIND (ON THIS PAGE) needs an exact match, so the text will not be selected when you view the page.
- Remember, no matter how carefully you compose your search to prevent ambiguous form, someone will have found a way to make that form ambiguous.

When I find a correct hit I usually copy the text around the phrase searched for and paste it into a Word document for archiving. I sometimes also copy the heading or name of the site, and the URL. I try to include a context, too, if it is not immediately clear.

When you have looked at the first 40 or 50 sites you will have a percentage of good hits to bad, and this can be applied back to the search figures to give a fairly accurate

frequency. For instance, if you find 36 useful pages out of 50, and you had an original page count of 15,600, then by calculation you have approximately 11,200 real hits. Of course, you also have 36 examples of the phrase you are investigating from a wide range of sources and authors, which is itself a useful mini-database.

Studies done using the Web

So what have I discovered so far by using the Web as a corpus? I will look at three studies I have done.

My first study was an analysis of the pragmatics of the phrases *I like me* and *I like myself*, which turned into a multi-dimensional case study. I looked at a range of emotive verbs (love, like, dislike, hate, etc), both singular and plural cases (me/us and myself/ourselves/ourself) and present and past tense (like and liked).

What I found was rather interesting. *I like me* was more common that *I like myself*, but this frequency reversed as different dimensions of the study were changed. So *we hated ourselves* was twenty five times more likely than *we hated us*. *I like me* was also used mainly as a stand-alone form related to self-help products, while *I like myself* was usually qualified (*I like myself at this weight*), only marginally associated with self-help products, but heavily implicated in the subject of personal weight. The difference between the internalised *me* and the externalised *myself* proved to be very important for my studies on language origins, an outcome I had not expected.

This study was only possible on the Web. I was dealing with phrase frequencies in the tens and hundreds, which would have been statistically insignificant on a smaller corpus.

A second study I made was on hendiadys, the relative frequency of *try to* and *try and*. This study was largely concerned with the verbs that collocated with the two phrases. I also tried to check for dialectic differences, but this was largely a failure; I could only conclude that both usages are common and no dialect appears to use only one of the forms. Comparing relative frequencies was impossible, and would require specialised corpora. The Web will never be able to replicate the detail that established corpora provide.

I did find some interesting things though: the infinitive form *to try and* was much more frequent than *to try to*; but the active form *try to* was much more frequent than *try and*.

I also discovered that there were marked differences in ratios depending on the verb used. Over 90% of cases with the verb *know* were *try to know*, but only 75% of cases for the verb *see* were *try to see*. There seems to be a difference between usage which is either semantic or phonetic – I think it's the latter, but I didn't prove it either way.

My third study involved the strange case of *came in out of*. This started as a study of triple prepositional forms, *in out of*, *down off of*, *up in by*, and so on. This got hijacked when I realised that 5% of *in out of* cases were *came* or *come in out of*, and

22% of those cases were *came* or *come in out of the cold*. Of course, there must be some contamination from *The Spy Who Came in from the Cold*, but why should *from* become *out of*? Indeed, why use *out of* at all? A look at the first 100 usages of *came in out of* found that thirteen used *the cold*, eight used *the rain*, and altogether 31 used weather nouns. There were nine cases of *nowhere*, eight of *the blue*, five of *order* and four of *shape*. This is a more complex study than I had believed, and is still ongoing. I still haven't started on any of the other possible triple prepositional forms.

Plagiarism

I will mention one final thing about the Web. A lot has been said about the capacity it provides for plagiarism. It is true that it is easy to find a text if you know how to search, and the lazy student can easily cut-and-paste an essay together. But the knowledgeable teacher can detect plagiarism just as quickly. **Appendix 2 – Spotting Plagiarism** provides an example of how to do this. The mechanisms that make it easy for the plagiarist can be turned against them, making counter-plagiarism just as simple.

Conclusion

So, I give you the wonder of our age, the World Wide Web: a curse for those who think linguistics should be hard work, a blessing for those who can make the effort to learn its ways. And for those who want to work with a massive corpus it is a godsend: a trillion words of unselfconsciously written data. Now surely you have enough words!

References

Helen Aristar Dry & Anthony Rodrigues Aristar, 1998. The Internet: an introduction. In John M Lawler & Helen Aristar Dry (eds), *Using Computers in Linguistics: a practical guide*. London, UK: Routledge.

Tim Berners-Lee, Wendy Hall, James Hendler, Nigel Shadbolt, Daniel J. Weitzner, 2006. Creating a Science of the Web. In *Science*, Vol 313, Issue 5788, 11 August 2006, pp769-771.

Celeste Biever, 2006. Google vs. Microsoft: The fight for your computer. In *New Scientist*, 20 May 2006, Issue 2552.

Martin Edwardes, 2003. I Like Both Myself and Me**.** In *CamLing 2003: proceedings of the University of Cambridge First Postgraduate Conference in Language Research*. Cambridge, UK: CILR.

Martin Edwardes, 2003. Trying Times. Unpublished paper available at: http://www.btinternet.com/~martin.edwardes/

Susan Hockey, 1998. Textual Databases. In John M Lawler & Helen Aristar Dry (eds), *Using Computers in Linguistics: a practical guide*. London, UK: Routledge.

Stig Johansson, 1991. Times Change, and so do Corpora. In Karin Aijmer & Bengt Altenberg (eds), *English Corpus Linguistics*. London, UK: Longman.

Michael Reilly, 2006. Internet Search Engines Go on Trial. In *New Scientist*, vol 191 Issue 2565, 19 August 2006, pp24-25.

James Simpson, 2002. Discourse and Synchronous Computer-Mediated Communication: uniting speaking and writing?. In Kristyan Spellman Miller & Paul Thompson(eds), *Unity and Diversity in Language Use*. London, UK: Continuum.

David R. Worlock, 2001. The best and worst of times. In *Nature* 413, 671-671 (18 Oct 2001).

Appendix 1 – Making Searches

Select your search engine. I recommend AltaVista or Google .
Make sure the search is set for Worldwide and not UK. See title bar on search screen.
Type in search keywords. There are three ways to do this:

- Type in *animal communication*. This will find all pages containing either the word *animal* or the word *communication*.
- Type in *"animal communication"* (including the quotes). This will find all pages that contain the exact phrase *animal communication*.
- Type in *animal + communication* (including the plus sign and spaces). This will find all pages that contain both the words *animal* and *communication*.

Search strings can be combined. So *"animal communication" + hauser + pdf* will find all pages containing the string *animal communication* and the separate words *hauser* and *pdf*. (pdf identifies links to Adobe Acrobat downloadable files).
If you get too many hits then add more keywords

Hints and Tips

Prepare for disappointment. There is an awful lot on the Web, but not much from before 1995 - although more is being added every day.
The more you know about what you are searching for the better. However, the more complex the search string the fewer "hits" you will get. This is fine if you are looking for a specific paper and find it, but you will miss many almost as interesting papers. Better to start the search with few keywords and add others if the hit rate is too high.
Searches do not include punctuation. So a search for *"he likes me"* will find *Those are the thinks he likes. Me, I prefer…* Letter cases (capitals and lower) are also ignored.
Sometimes a search will identify the end of a keyword as the end of a word, sometimes it won't. So a search for *"I like me"* could find *ferengi like meat*. This can be overcome by putting a space before and after the string: *" I like me "*.
When searching for names, remember that there are various conventions available: *Martin Edwardes*, *M. Edwardes*, *Edwardes et al…* It is best to search for *edwardes* and try to cut the search down in some other way, e.g. *edwardes + linguistics*.
When looking for something specific, set a time limit for surfing (half an hour is usually enough). If you find something useful then save it to disk and move on.
When you find an interesting page (not an acrobat document) and want to save it:

- Click on Edit then on Select all. Click on Edit again then Copy.
- Open Word, or open a new document if Word is already open.
- Click on Edit then on Paste.
- Put the page address onto the top of the document or as a header on each page.
- Save and close the document.

When you find an interesting Acrobat document and want to save it: Click on the Save symbol (the little disk picture) on the acrobat tool bar. Locate an appropriate folder; give the file a name; press Save.

A Sample Session

I want to find a copy of W. T. Fitch's article on the descent of the larynx from one of the evolution of languages conferences, but I can't remember which one. I can use the following procedure:

- Search for *fitch*. This gives 218,966 results: too many.
- Search for *fitch + larynx*. This gives 178 results: better, but still a lot to trawl through.
- Search for *fitch + larynx + evolution*. This gives 49 results: close enough.
- Look down the list. Find a pdf in the list labelled **FitchWray**. Check it out: Bingo!

Appendix 2 – Spotting Plagiarism

You suspect that the following piece of text in a student's essay has been plagiarized. It does not fit the style, or it just seems out of place:

> Consider as well a native Russian living in America who has been speaking English for years and is fully conversant. The Russian may have heard Americans use definite articles a myriad of times and yet the Russian says, "I have key to apartment." and he would feel this to be grammatical, since the Russian language does not use articles, as English does. (The Russian feels no more need to say "the key" or "the apartment" than we do to say "the France", although if we were speaking French we would have to add the article to be correct.) Since we can easily understand what the Russian means, and since he feels that he is expressing all the necessary ideas without the use of articles, then, again by the Chomskyan methodology, the Russian must be speaking grammatical English.

Go to a Web search engine (I have used AltaVista) and type in a selection of text from the essay – between five and ten words is about right. I have used "since the Russian language does not use articles". Enclose the text selection in double quotes to ensure exact matches only are selected. You will get a list somewhat like the following:

> **AltaVista found 6 results**
>
> FrontPage magazine.com :: Chomsky's Linguistics Refuted by John Williamson
> ... and he would feel this to be grammatical, **since the Russian language does not use articles**, as English does ...
> www.frontpagemag.com/Articles/ReadArticle.asp?ID=16508
> More pages from frontpagemag.com
>
> Chomsky's Linguistics Refuted
> ... and he would feel this to be grammatical, **since the Russian language does not use articles**, as English does ...
> www.frontpagemag.com/Articles/Printable.asp?ID=16508
> More pages from frontpagemag.com
>
> News
> ... and he would feel this to be grammatical, **since the Russian language does not use articles**, as English does ...
> www.africancrisis.org/NewsView.asp?Rec=4119&Action=V&Sort=D&Page=1
> More pages from africancrisis.org
>
> Chomsky's Linguistics Refuted
> ... and he would feelthis to be grammatical, **since the Russian language does not use articles**, asEnglish does ...
> www.discoverthenetwork.org/Articles/z-Chomsky.htm
> More pages from discoverthenetwork.org
>
> LP: Chomsky's Linguistics Refuted
> ... and he would feel this to be grammatical, **since the Russian language does not use articles**, as English does ...
> www.libertypost.org/cgi-bin/readart.cgi?ArtNum=115069
> More pages from libertypost.org

Select one of the links and you will see the plagiarized text in full.

Appendix 3 – Glossary

Domain name – the name by which your site is known to the Web, e.g ***www.cityacademy.co.uk***. It is sometimes show preceded by http://, but this is now unnecessary.

HTML (HyperText Mark-up Language) – the particular form of Hypertext used on most of the Web. It has been extended as XTML and XML, but the basic form of HTML remains the backbone of most web resources.

Hypertext – this is essentially what Tim Berners-Lee invented. It is a universal coding format that allows pages to be produced and displayed in a standard format. It requires two components: a program to read the text and translate it into screen images (Explorer, Netscape and a few other tools do this); and a program to write the text (this can be done with Notepad if necessary, although FrontPage, Dreamweaver and several other design tools are available to make the coding very simple. You can even produce web pages in Word if you wish).

Link – a line of text or picture on a web page which makes another web page load. Different sites show links in different ways, but if the cursor arrow changes to a pointing hand when moved over the text or image, it is a link.

Internet – a set of web pages which link together to produce a massive free-form database. This is often seen as synonymous with the World Wide Web, but the Web is the set of indexed pages on the public Internet. It is possible to create pages that are not indexed to the Web and are invisible to search engines.

Intranet – a limited, private internet. The pages are held on a server which is not registered to the Web, so can only be viewed if you have access to that server.

Page – the basic unit on the web. Resources are accessed via pages, and a related group of pages makes a site. The page itself can be any size.

Resource – a picture, video, sound file, document, etc. Usually these are the bread-and-butter of the Web, but it is the Web-site-page-resource hierarchy which makes them into a database.

Search engine – a program which gives access to a massive searchable index of the Web. Google has 50% of the market, Yahoo nearly 25%, and the rest share what's left. Google remains the best one for linguistics searches, although I tend to use AltaVista, which is almost as good.

Service Provider – a company which offers webspace and domain names for a price. Universities sometimes provide this service to their staff for free.

Site – a set of pages which have been set up to work as a coherent unit. Often the pages are physically located on a single server, but this is not actually necessary. The logical map of the Web is becoming very different to the physical map.

URL (Uniform Resource Locator) – the name of a page on the Web. Usually it consists of the domain name plus an extension, e.g. ***www.cityacademy.co.uk/about.html***.

Web – any collection of pages. Used in this paper as a shortened form of World Wide Web.

World Wide Web – the vision of Tim Berners-Lee. A set of pages containing information in various formats (visual, auditory and written) which are universally available. It is therefore a giant database which, theoretically, could hold the World's knowledge in public form.