San José State University School of Library & Information Science

LIBR 244 Online Searching (Tucker) – Spring 2005

Michael Stoler – Final Essay – 15 May 2005

Online Resources for Lexicographical Searching

ABSTRACT

Professional online searchers may for a variety of reasons be asked to determine the earliest use of word or other aspects of its meaning and history. There are three main kinds of resources for this: large fulltext corpora, general search engines, and compiled dictionaries. Each has advantages and disadvantages.

Imagine that you are a professional searcher, and someone comes to you needing information about the origin of a word or phrase, its meaning and usage and how it developed over time, and/or when it came to be used with a particular meaning. These sorts of questions could come up in several environments. At a news organization, editors might have noticed a term being used in public discourse, and want to know if it has special significance to those using it, or if for the organization's staff to use it would suggest some bias. (Malesky (2005)) Public relations firms, advertisers, marketers, and naming and branding companies, considering new names and slogans for commercial products and services, might want to check for possible negative associations with the words they plan to use, and if they hope to trademark them, will want to verify that the words have not become "generic" or "descriptive", and thus ineligible to be protected. Many other legal matters can hinge upon the meaning of a word, as cases are argued based not just on what a law or contract requires, but on what the terms actually mean, or meant at the time of original drafting; if there are no definitions within the document text, it will be necessary to refer to outside sources for them. For instance, a Federal law imposes harsher sentences for drug possession on those who "carry" firearms; a 1998 case tested whether the verb here meant to bear on one's person, or to transport in a vehicle, as the defendant had done. (U.S. Supreme Court (1998))

Some of these questions can be answered by reference to a good hardcopy dictionary found in most homes or offices. But others require more extensive research, requiring resources found only in libraries (and large ones at that.) What this essay aims to address are the options for someone sitting in front of a computer with Internet access. I have looked for information about the major types of online resources: general-purpose

search engines, online dictionaries, and large databases and corpora. I have tested most of them myself to determine the extent, temporal reach, and currency of their holdings and information returned, using a variety of test words, including recent ones such as “blog”, “phat”, and “d’oh” (the Simpsonian interjection), words and phrases with longer histories but recent specific uses, such as “weapons of mass destruction”, “pledge drive”, “filibuster”, and “ethnic cleansing”, and long established words like “mortgage” and “carry”. Since it describes the online world, whose vastness and rapid evolution are proverbial, this essay will not attempt to be exhaustive or scientific, but rather to present representative examples of online resources, and discuss their use and their uses, their advantages and disadvantages, and come to some general conclusions.

The gold standard for online word searching has for some time been Nexis. (Shapiro (1986), 143; (1998), 280-281) Its use for “antedating” (finding earlier citations than the ones on record) was pioneered in the 70’s and 80’s by Fred R. Shapiro, a librarian at Yale University and Lecturer on legal research at its Law School. (Shapiro (1984)) He originally used legal databases such as a Westlaw and Lexis, figuring that legal language, while different from everyday English, could still give a good reflection of it. (Shapiro (1984)) William Safire, writer of the “On Language” column in The New York Times, makes frequent use of Nexis searches, and phrases such as “a Nexis search reveals” have hundreds of hits in Google.

Looking for information about words themselves, one avoids the major difficulty of online searching: finding material on a particular subject based only on the words used to describe it, since in this case, the word *is* the subject. To find out how a word is being used, you simply search for it in fulltext, then browse the results, or combine the word

with others in order to test whether it is being used in certain contexts. (Shapiro (1986)) With billions of words of text, covering a wide variety of publications, in dated articles which can be shown in historical order, all searchable from one powerful input form, Nexis is an easy and convenient (if expensive) way to find citations illustrating a word's development.. The size and representativeness of the database even permit comparisons of numbers of hits in various contexts, for statistical studies. (Drum (2005), Nunberg (2004))

Nexis does have a few peculiarities and drawbacks. It is still really oriented towards subject searching, so that if your search phrase turns out to be a subject heading, like "weapons of mass destruction", Nexis will return every article on that subject, whether it contains the exact phrase or not. (Use the "body" specifier before the terms to avoid this.) Articles are displayed newest first, requiring several clicks to get to the earliest citations. It is easy to get more than the 1000 citations that the search engine can display, requiring you to limit your search temporally to get a displayable set (though this process of "walking the search forward" can be interesting for a term such as "ethnic cleansing", as its used increases exponentially.) Also, some articles are misdated, producing confusing results. Shapiro (1998, 1986) notes a deeper problem in searching, that of words that have multiple meanings (such as "crack" (Simpson (1986)); that have homographs, other words with the same spellings; or that are so common that they retrieve "junk citations". (Barnhart (1985)) These writers also note that Nexis is limited to the sort of written English found in newspapers, and thus may give little evidence about the origins of colloquialisms. Landau writes: "The billions of words in Nexis and other databases represent a mere drop of water compared to the ocean of discourse that

occurs daily, and statistical data of frequency, sense, or usage can only measure the items included within their survey, not the English language." Perhaps most important, Nexis only covers from the mid-70's forward. For earlier word history, one must seek other sources.

Dialog has even more limited potential for word origin searches. It actually is possible to order the results of a search by date to find the earliest citation, using SORT and PD, but the biggest problem is the lack of a single large fulltext file equivalent to Nexis's AllNews. Though they cover many of the same sources, File 20, Dialog Global Reporter, goes back only to 1997, as does File 781, ProQuest Newsstand™, while File 469, Gale Database of Publications and Broadcast Media, only covers the current year. Individual newspaper files go back to the late 1980's. The same applies to DialogWeb; the largest assemblage of newspapers can only be searched by subjects, not words in the text. Some of the power of Nexis can be approximated on DialogNews, but without a KWIC format option, there is no way to avoid scanning through an entire retrieved article to find one's searched words, and the lack of proximity operators makes precise searching difficult.

Dialog's wide range of specialized files can be useful, however, in learning how words are employed in discussing specific subjects (these narrower contexts can also reduce the problem of polysemy.) And abstracts can show as wide a variety of vocabulary as fulltext. Another highly regarded database of academic and scholarly, rather than popular and general material is JSTOR. Shapiro (1998) notes that some of the journal archives it contains go back as far as the beginning of the 20th century, much further than Nexis, and it can all be searched from a single screen. In fact, there are

available online many other "corpora," large assemblage of documents supporting full-text searching, but organized (and often tagged) in such a way as to permit the ready and precise identification of citations by date and author. (Nunberg (2004) differentiates these "occurrent corpora" from others containing random samples of language, designed for the sorts statistical studies on word frequencies, grammatical structures, etc., undertaken by highly trained linguists, and unlikely to be asked of general searchers.) Many corpora have specific time restrictions, such as Old or Early Modern English; input through enormous effort, they extend the reach of electronic lexicography far back through the centuries. Unfortunately, they tend to be limited in size compared to Nexis; both the American National Corpus (www.americannationalcorpus.org), and its model, the British National Corpus, contain one hundred million words. (Emmons) The Making of America (MoA) corpus (http://www.hti.umich.edu/m/moagrp/) is one of the most impressive, containing over 10,000 books and 50,000 articles from 1800 through 1928, all searchable (using Boolean, and proximity) from one input box. With a few clicks, I was able to view examples of the use of my terms in their original contexts (PDFs of the source books.)

Perhaps the simplest solution to a lexical search problem is that Swiss Army knife of search engines, Google, and in many ways it is not a bad one, used by professional lexicographers like the Oxford American Dictionary's Erin McKean. (Montagne (2005)) In effect, the entire Web becomes the fulltext database. Google's vast reach, its speed, and the Advanced Search capabilities, allowing specification of combinations of words, and limitation by language and dates of pages, mean that you can get a general idea of the use of a word very quickly, and make statistical comparisons based on numbers of hits (see above, re: "Nexis search". Of course, as with Nexis, if you cannot come up with a

way to specify a meaning of a word contextually, you have to wade through all the citations and classify them by meaning yourself.)

But using Google has its limitations as well, similar to those other sorts of word searching: junk citations, polysemy, ambiguity. Although Google indexes a much wider variety of content than Nexis, this can be a disadvantage; even limiting searches to ".edu" sites will not screen out directories, lists, and other uses of words that do not represent speech or writing. Google's reach is limited, as Mary Ellen Bates (2004)has discussed, by passwords, search screens, etc., though it could be argued that it still gives a good cross section of the language in use. Using the Web in general as a source of citations can be problematic, due to the difficulty of dating webpages. (In fact, the Oxford English Dictionary avoids Web references for this reason. (OED (2005) – "Documentation")) The original content of the Web only covers the last ten years or so, though of course there is much older material that has been put up there, but then one has to wonder about the accuracy of the transcribing. (Some archives of newsgroups go back a lot further, and tend to be dated quite precisely, however.) Google News, recommended by McKean (Montagne (2005)), has better date ordering, but even less time reach.

The biggest problem with searching directly on the Web or on corpora, though, is that it requires the searcher to make linguistic judgments him or herself. While this may seem like an ability most people have as speakers of the language, in some cases it may require expert knowledge. Thus, just as people turn to expert-produced directories instead of search engines, they may prefer expert-assembled dictionaries to corpus searching. Barnhart estimates that there are 12,000 new words or meanings of words arising each year, yet Erin McKean (Saroyan (2005)) estimates that only100 to 500 of them enter the

dictionary. And all this compilation takes time, meaning that dictionaries will not be able to explain every new word turned up in a text search, but will explain better those that they select.

Most of the major dictionaries have online versions, but with limitations. Some simply put the most recent edition of the hardcopy work online, with no attempt to keep it current; the American Heritage Dictionary (accessible through www.bartleby.com) is the 2000 edition (though of course, it has the terrific Indo-European etymologies edited by Calvert Watkins of Harvard); Cambridge University Press's website (dictionary.cambridge.org) gives access to the Cambridge Advanced Learner's Dictionary and other relatively small works as a way of promoting their sale. Others only allow access to part of the contents, and require a subscription to view the full information (Merriam-Webster (www.m-w.com), offers an online version, but charges $29.95 to view its 470,000 entry unabridged one.) None of these feature dated the dated citations key to historical lexicography.

There are also several sites that access a variety of dictionaries, sometimes by a single search. Dictionary.com uses Webster's New Millennium™ Dictionary of English, Preview Edition (v 0.9.6), American Heritage, and Princeton University's WordNet (wordnet.princeton.edu, which is more a list of synonyms, or a thesaurus, than a dictionary), among others, and gives some dates, though only limited citations. Dict.org uses WordNet to supplement what it calls "The Collaborative International Dictionary of English v.0.48", which, it notes, is "derived from the Webster's Revised Unabridged Dictionary Version published 1913" (!), and OneLook (www.onelook.com/?d=all_gen) searches a truly dazzling array of different works.

The problem of the lack of dated citations to illustrate past meanings can to some extent be solved by consulting older dictionaries. MoA contains several important historical dictionaries; Yourdictionary.com links to Noah Webster's 1828 dictionary, among others (http://yourdictionary.com/languages/germanic.html#english); and Project Gutenberg (www.promo.net/pg) contains this and others for downloading. (Unlike MoA, Project Gutenberg cannot be searched as a whole, so it is not a corpus.) Unfortunately, Samuel Johnson's 18th century classic is not online, nor is H.L. Mencken's "The American Language", nor the Dictionary of American Regional English. Furthermore, older dictionaries may not have been compiled according to the same standards as modern ones, for instance, using only citations from literary sources rather than popular ones, or omitting terms which seemed too scientifically specialized. Some must be downloaded and searched with special programs, rather than through a Web browser.

There is a plethora of smaller dictionary sites, assembled by individuals with an interest, or by the efforts of large numbers of people who send in entries. (One strategy for finding them is simply to type the target word, and "definition" or "etymology", into Google.) As may be expected, they vary enormously in quality, and as each one tends to cover only a few thousand words that strike the authors' fancy, the chances of finding reliable information on the word you want can be small. The collaborative Wikipedia has an associated "Wiktionary" (www.wiktionary.org), with about 70,000 entries, but they are of irregular quality, and without dated citations. Douglas Harper's "Online Etymological Dictionary" (www.etymonline.com) contains a large number of word histories, with dates, largely taken from the OED and the Barnhart Dictionary of Etymology. (Harper (2001)) Paul McFedries' Wordspy (www.wordspy.com) has an

earliest citation listed for every entry, but tends only to feature the trendiest words. Some sites are more like periodicals than reference books, with articles on specific words and phrases; their archives, however can be a valuable resource. (The American Dialect Society website (www.americandialect.org) contains the last 15 years of the journal American Speech, e.g.)

The grandmother of all dictionaries is the Oxford English Dictionary (www.oed.com). It can be accessed online through major libraries; an individual subscription costs $295 a year. (A free but limited version is available at www.askoxford.com. The bound version costs $1500; the CD-ROM, $295. And the publisher claims not to be turning a profit! (OED (2005) – About)) The Online OED is descended from the original Oxford English Dictionary, conceived in 1857, begun in 1879, finished in 1928, and supplemented thereafter. In 1984, the editors embarked on a five-year project to digitize the dictionary and its files of 2.5 million citations, leading to the publication of the Second Edition in 1989 (on CD-ROM in 1992.) For the Third Edition, all entries are being rewritten, and, unlike, for instance, American Heritage, the updates are added to the online version every three months, marked as "draft" with the date of addition. Until December of 2004, the new edition and the Second Edition were two separate files, but they have now been integrated, with the new entries replacing the old ones, though the latter are still visible by clicking a button to launch a pop-up window.

The Online OED is a joy to use. One can look up a specific word, or search for words or phrases or combinations, in the entries or the citations. The advanced search function allows for proximity searching, and restriction to various parts of the entry or

parts of speech, and all one's searches are remembered for later reference. The main screen shows the entry, its meanings, and the citations for it, while a sidebar at the left shows where the word falls in the alphabetical list (allowing the user to quickly identify other forms) or an outline of the meanings and subheadings for the entry (useful for looking through the many meanings of words like "carry".) A timeline shows where the dates of the citations fall. The coverage is complete, going back almost a thousand years earlier than Nexis, and yet up-to-date, though some currently trendy words turn out to have longer histories. ("D'oh" goes back to 1945.)

On the other hand, it is not perfect. The "definitive record of the English language" (OED (2005)) contains typos (I found one in a citation from Tom Wolfe for "push the outside of the envelope".) The citations are chosen to mark key points in the semantic evolution, rather than giving any statistical indication of the preponderant meaning, and they are all you see; associated information allows you to identify the works whence they come, but it would be up to you to find the actual text, so that you cannot judge the larger context in which the word appears.

Then there is the question of the OED's sources. To gather citations, it relies on "reading programmes", having volunteers monitor specified (but wide) swathes of literature for new words and usages. On Shapiro's urging (Shapiro (1984)), the editors overcame their resistance to using electronic sources at all (Burchfield (1980)), but even today they employ databases such as Medline, JSTOR, Dialog, Nexis, and even Melvyl, (the University of California's union catalog, whose millions of entries represent a huge tranche of text), for verification once an item is identified, rather than primary research. For some uses, the OED might just not be hip enough.

Perhaps the biggest problem in online word searching is the nature of the question. No search can be entirely exhaustive. Barring a transcription or digitization error, not an uncommon occurrence, a citation from a certain date shows that the word was used in that way at that time by that one person, but not necessarily commonly. The lack of a citation does not mean the word was not used, just that we cannot find a citation. In 1984, at the dawn of the age of online searching, Shapiro wrote: “The historical lexicographer’s dream, a computer terminal that instantaneously displays the first appearance in print of any word, word combination, or phrase, will never fully be realized, but future scholars will benefit from full-text data bases that can retrieve the earliest usage of a given term in the documents within their coverage.” Using Nexis, historical corpora, Google, and the OED, we are close to fulfilling his hope and prediction – if we can agree on what the information means.

References:

Barnhart, David. (1985) Dictionaries: Journal of the Dictionary Society of North America, 7, p253-260

Bates, Mary Ellen (2004) "Now that You've Fired Your Boss", Searcher, May, p8-15

Burchfield, Robert W. (1980) "Aspects of Short-Term Historical Lexicography" Proceedings of the Second International Round Table Conference on Historical Lexicography. Dordrecht, Netherlands: Foris. p271-286

Drum, Kevin (2005) "Political Animal", Washington Monthly, April, accessed at www.washingtonmonthly.com/archives/individual/205_04/006182.php , on May 15, 2005

Emmons, Kimberly (date unknown) "Electronic Resources for English Language Research" accessed at http://home.cwru/edu/~kke1/engl310/resources.htm , on May 15, 2005

Harper, Douglas (2001) "Introduction & Abbreviations", accessed at http://www.etymonline.com/abbr.php , May 15, 2005

"JSTOR: No. 4, Issue 3, JSTORNEWS, November 2000", accessed at http://www.jstor.org/news/2000.11/words.link.html , on May 15, 2005

Landau, Sidney I. (1989) "Of Lexicography, Computers, and Norms" American Speech, 64:2, p162-163

Logan, Harry M. (1991) "Electronic Lexicography" Computers and the Humanities, 25: 6, p351 - 361

Malesky, Kee (2005) Telephone interview with librarian at National Public Radio, conducted by Michael Stoler, March 3

Montagne, Renee (2005) "Googling for New Words" Morning Edition (National Public Radio (NPR), interview with Erin McKean), April 18, accessed at http://www.npr.org/templates/story/story.php?storyId=4604623

Nunberg, Geoff (2004) "Cyber-Infrastructure for Linguistic and Language Study", accessed at http://www.acls.org/cyberinfrastructure/ACLSSlides.pdf , on May 15, 2005

Oxford English Dictionary (2005) "About the Oxford English Dictionary" (oed.com/about), "History of the Dictionary" (oed.com/about/history.html), "Documentation - Preface to the Third Edition" (oed.com/about/oed3-preface), "Writing the OED – Electronic Resources" (oed.com/about/writing/resources.html), "*Oxford English Dictionary* News" (oed.com/news), accessed on May 15, 2005

Saroyan, Strawberry (2005) “In Land of Lexicons, Having the Last Word” The New York Times, CLIV: 53,158 (March 19), pB9

Shapiro, Fred R. (1984) “A Computer Search for the Origin of *Executive Privilege*” American Speech, 59:1, 1984, p60-62

Shapiro, Fred R. (1986) “Yuppies, Yumpies, Yaps, and Computer Assisted Lexicology” American Speech, 61:2, p139-146

Shapiro, Fred R. (1998) “A Study in Computer-Assisted Lexicology: Evidence on the Emergence of *Hopefully* as a Sentence Adverb from the JSTOR Journal Archive and Other Electronic Resources” American Speech, 73:3, p279-296

Simpson, John (1986) “Computers and the New OED’s New Words” Euralex International Congress, Zurich, p437-446

U.S. Supreme Court (1998) “Syllabus v. Certiorari to the United States Court of Appeals for the Fifth Circuit , No. 96-1654. Argued March 23, 1998 - Decided June 8, 1998” accessed at http://66.102.7.104/search?q=cache:3VVmwawjSTcJ:caselaw.findlaw.com/scripts/getcase.pl , on May 15, 2005