Concordancing the Web:

Promise and Problems, Tools and Techniques[1]

William H. Fletcher
fletcher AT usna DOT edu
United States Naval Academy

Submitted for publication in:

Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer, (eds.). *Corpus Linguistics and the Web.* Amsterdam: Rodopi. (2005?)

http://kwicfinder.com/FletcherConcordancingWeb2005.pdf

Abstract

The Web is an inexhaustible reservoir of machine-readable texts in most of the world's written languages, available for compiling corpora or consulting directly as a "corpus". This paper first surveys some characteristics of the Web and discusses the potential rewards and practical limitations of exploiting the Web either directly as a linguistic corpus or to compile corpora. Particular attention is paid to search engines, our gateways to the Web. Next the author reviews several innovative applications of Web data to corpus-related issues. KWiCFinder (KF), his application to help realize the Web's promise for language scholars and learners, is described and motivated in detail. Readily accessible to novices yet powerful enough for advanced researchers, KF conducts Web searches, retrieves matching online documents and produces interactive keyword in context concordances of search terms. He details KF's enhancements to AltaVista search and the limitations imposed by working with market-driven search engines. This paper then discusses the pitfalls of "webidence" in serious research and proposes an initial solution. Finally the author reviews the future of the Web for corpus research and application.

1. The Nature of the Web

1.1 Size, Composition and Evolution

The World Wide Web is a wondrous place, with an overwhelming range of languages, content domains and media formats. Just how many webpages there are and how they are distributed by language and genre are not easy questions to

---

[1] Portions are based on an earlier paper on pedagogical applications of Web as corpus available online (Fletcher 2001). It was substantially revised and updated in spring 2005 during a sabbatical at the Radboud University of Nijmegen. The author gratefully acknowledges the RU Language and Speech research group's generous hospitality and the Naval Academy Research Council's partial support of this research.

answer. The Web is constantly changing and growing, and even the best estimates can only approximate its extent and composition. Studies of the nature of the Web echo the story of the blind men and the elephant: each one extrapolates from its own samples of this ever-evolving entity taken at different times and by divergent means. The most reliable estimates suggest that the number of publicly-indexable webpages in mid-2005 falls in the range of 10 to 20 billion (i.e. thousand million; see e.g. Gulli and Signorini 2005); some speculate that the actual number is far greater.

These ten billion-plus easily accessible pages are only the tip of the iceberg. To be indexable, a page must allow unrestricted public access, and another publicly accessible page must link to it with a standard HTML tag.[2] Far larger is the vast "invisible" Web of content in databases, which cannot be "crawled" (explored) by an all-purpose "robot" (crawler program), only explored by entering relevant queries in a form.[3]

How dynamic and volatile the Webscape has become is revealed in an exhaustive year-long study of 154 websites from the perspective of a search engine (SE) (Ntoulas et al. 2004). This selection of commercial, government, academic and media sites primarily from the U.S. was judged "representative" and "interesting", with content that would rank high in a link popularity scheme like Google's PageRank (see below). The authors' software monitored these sites weekly, averaging 4.4 million webpages (65 GB) per crawl. From their analysis of these pages over time, they estimate that new pages[4] appear at the rate of 8% per week. Assuming 4 billion total webpages at the time, they extrapolate their figures to 320 million new pages, or roughly 3.8 terabytes of new data for the Web as a whole each week. Here "new" does not mean "additional" or even "novel": the total number and size of webpages on these sites stayed relatively constant as old pages retired, and only about 5% of the weekly harvest actually represented new content; 50% of the online content remained available a year later. Far greater volatility was documented in the link structure: each week 25% new links were created, and only 20% of links survived a year. Underscoring the importance of sites like the Web Archives' Wayback Machine,[5] the authors speculate from other evidence that only

---

[2] *Indexable* is distinct from *publicly accessible:* search engines (SEs) "crawl" the Web by following links from known sites to pages not yet in their database, downloading then extracting links from these new pages and following those new links. If a site has no incoming link from sites known to the crawler (and it is not submitted for indexing), its pages will never be found, so they are not indexable even if publicly accessible. Wikipedia gives an overview of SEs and crawling, and Chakrabarti (2002) demystifies and details their workings.

[3] Ntoulas et al. (2005) propose a framework for generating queries to crawl deep-Web sites which succeeded in downloading up to 90% of "hidden" content of generally high quality.

[4] From the SE perspective of their paper the authors count *any* changed page as a new page, even if only a single word or the URL changes.

[5] The Web Archive (http://web.archive.org) preserves over 40 billion webpages from 1996 on for public access. Not a comprehensive repository, it retains pages sampled at different times from archived websites, affording a glimpse into their evolution over this period. The Archive also helps preserve and distribute audio, video and print materials. In a study of papers in three online scholarly journals, Ho (2005) found less than 50% of links functional. Archives dedicated preserving access to papers in mathematics and the natural and information sciences such as arXiv and CiteSeer deserve emulation by other disciplines as well.

20% of all webpages are still accessible a year later. While this investigation addresses characteristics of the bulk of the Web in deep websites, not its breadth, it graphically portrays the rapid radical evolution of the Net.

While such establishment websites remain stable in size, there seems no end in sight for the colossal growth in number and sheer text volume of self-published and collaborative websites like blogs (Web logs) and wikis, which often feature thoughtful, well-written content. During the first half of 2005, blog articles indexed by Bloglines.com doubled to over 500 million, and Blogwise.com lists blogs from 190 different countries.

1.2 Languages on the Web

Despite the Web's overwhelming size and global expansion, English continues to predominate. Studies by Inktomi and Cyveillance (Moore and Murray 2000) in 2000 conclude that at that time over 85% of publicly-accessible webpages were in English. Around the same time, the *Fifth Study of Language and the Internet* (FUNREDES 2001) documents strong growth among the non-English languages in the proportion of webpages relative to English and observes that the number of webpages in the Romance languages and German was roughly proportional to the population of Web users with those languages as native tongue. O'Neill et al. (2003) find that the English-language share of the Web had dropped to 72% by 2002. In sharp contrast to the Web's first decade, recent years have seen no systematic studies based on large-scale general sampling of actual webpages. This hiatus presumably stems from the tremendous resources required and the limited (and brief) validity of any results. Nevertheless, current data from the principal SEs provide a rough indication of the webpage distribution by language. They suggest that English-language documents comprise around two-thirds of the content indexed in early 2005.[6] The large international SEs' bias toward the major European tongues, especially English, probably inflate their position relative to minority and non-Western languages in these data.

Historically Anglophone users and content have overshadowed other languages on the Net, but the trend toward diversity is clear and growing. Statistics compiled by Global Reach illustrate this long-term development. In 1996, four-fifths of the 50 million Internet users were native speakers of English. By September 2004, Anglophones constituted only 35% of the world's estimated online population of 801

[6] Currently Google, Yahoo, MSN, and Teoma are the only large (over two billion pages), well-established independent SEs; others either are smaller or use other providers' databases. A technique outlined by Mikami (2003) was adapted to estimate the proportion of pages in English: I conducted a series of searches on the first three of these SEs for common digits *(1, 2, 3, 2003, 2004, 2005)* on webpages in English, then in all languages, and calculated the proportion between the two tallies. MSN also supplies tallies of pages *without* these digits. For example, the sum of hits for the queries *2005* and *–(2005)* (i.e. with and without *2005* respectively) approximates MSN's total page count for all pages in all languages. The proportion of English pages for these queries fell in the range 60-70%. Grefenstette and Nioche (2000) offer a methodologically interesting study estimating the number of words (not webpages) online in various European languages, updated by Kilgarriff and Grefenstette (2003).

million.[7] Currently the Language Observatory Project and its Cyber Census Survey aim to raise awareness of the digital divide between languages and writing systems and track the distribution of languages online (Mikami and Suzuki 2005), and UNESCO is actively promoting linguistic and cultural diversity on the Web. The phenomenal growth in the non-Anglophone segment of the Web is spurring expansion of online resources in other tongues, particularly the smaller non-Western ones, to the benefit of those who investigate, teach and learn these languages.

2. The Web as a Corpus for Investigating and Learning Languages

2.1 Why Use the Web as Corpus?

The abundant online texts both tantalize and challenge linguists and other language professionals: the Web's self-renewing machine-readable body of documents in scores of languages is easy to access, but difficult to evaluate and exploit efficiently. Yet there are powerful reasons to supplement existing corpora or create new ones with online materials.

- **Freshness and spontaneity:** the content of compiled corpora ages quickly, while texts on contemporary issues and authentic examples of current, non-standard, or emerging language usage thrive online.
- **Completeness and scope:** existing corpora may lack a text genre or content domain of interest, or else may not provide sufficient examples of an expression or construction easily located online; some very productive contemporary genres (blogs, wikis, discussion forums…) exist only on the Net.
- **Linguistic diversity:** languages and language varieties for which no corpora have been compiled are found online.
- **Cost and convenience:** the Web is virtually free, and desktop computers to retrieve and process webpages are available to researchers and students alike.
- **Representativeness:** as the proportion of information, communication and entertainment delivered via the Net grows, language on and of the Web increasingly reflects and enriches our tongue.

2.2 Corpus Approaches to the Web

The term "Web corpus" has been used for at least three distinct concepts: a static corpus with a Web interface, one compiled from webpages, and the body of freely available online documents accessed directly as a corpus. We will disregard the first

[7] Data from Global Reach (2004), whose archives from September 2004 on the size the global online population tabulate the following percentages of users for the top ten languages: English 35.2%, Chinese 13.7%, Japanese 8.4%, Spanish 9.0%, German 6.9%, French 4.2%, Korean 3.9%, Italian 3.8%, Portuguese 3.1%, Dutch 1.7%, Other 10.1%. While Global Reach no longer updates these estimates, its data tracking the developments over the period 1996-2004 and explanation of the methodology are invaluable. Other studies of online populations include "Internet Usage by Language Statistics" (http://www.internetworldstats.com/) and the Computer Industry Almanac's forecast of one billion users online in 2005 (http://www.c-i-a.com/pr0904.htm).

sense and, following De Schryver (2002), distinguish between "Web *for* Corpus" (WfC), as a source of machine-readable texts for corpus compilation, and "Web *as* Corpus" (WaC) consulted directly. A well-known descriptive framework for finding and using information distinguishes three basic approaches: *hunting,* or searching directly for specific information, *grazing,* or using ready-made data sets composed and maintained by an information provider, and *browsing,* or coming across useful information by chance (Hawkins 1996). Each approach can serve as a model for corpus building or utilization. In the following sampler of applications of Wf/aC we use the *hunting* metaphor for SE-mediated access to the Web and *grazing* for systematic data collection on sites predetermined to be productive.

2.2.1 Hunting

Since the dawn of Web civilization, Anne Salzman and Doug Mills (2005) have sent their ESL students on "Grammar Safaris". Guided by their online assignments and armed only with a browser and a SE, they hunt down webpages with the structures they are studying, then find examples within the documents and copy and paste them into a word processor document to bring to class for discussion. In a comparable approach, Robb (2003) outlines browser-based techniques for researching English usage with Google.

WaC for language learners can be far more sophisticated than such Info-Stone-Age safaris. The Lexware Culler (Dura 2004)[8] enhances Google search with wildcards, part-of-speech variables and automatically generated morphological variants. It retrieves search engine report pages (SERPs) and displays only the snippets (the 10-20 word document extracts on SERPs) which match the user's potentially more specific query. While snippets may be too brief for some purposes and only a few languages are fully supported, Lexware Culler is a powerful proof-of-concept for WfC. One desktop application, WebLEAP (Web Language Evaluation Assistant Program), even automates the search phase for non-native English writers (Yamanoue et al. 2004). As they enter text, it displays Google SERP snippets of keywords to suggest appropriate wordings. WebLEAP also helps them judge text quality by displaying Google's hit counts of sub-sequences of their writings: rare or missing phrases are likely suspect. Chesñevar and Maguitman (2004) have proposed a comparable but more sophisticated solution yet to be implemented. Finally, Squirrel, a metasearch engine in development, promises to help locate suitable texts for language instruction and practice through automatic document classification and metrics of text difficulty and similarity (Nilsson and Borin 2002; Carlson et al. 2005).

Linguistic researchers also follow the hunting model to exploit the Web. To compile a dictionary of regional variants of German, investigators trawled the Web to complement the traditionally-compiled corpus materials gleaned from other sources

[8] http://82.182.103.45/lexware/concord/culler.html

(Bickel 2000, 2004).[9] Another study contrasts slogans from the 80s and the 00s as metaphors for their respective times; the former survive only in precompiled corpora, while the latter had to be studied via WaC (Gerbig and Buchtmann 2003). Other innovative solutions based on Web searching techniques include using the Web to disambiguate natural language confusion sets (Banko and Brill 2001), as a resource for example-based machine translation (Grefenstette 1999; Way and Gough 2003), to identify and collect sets of morphologically related lexemes (Tanguy and Mathout 2002), and to estimate frequencies of bigrams unattested in a given corpus (Keller and Laplata 2003). Kilgarriff and Grefenstette (2003) summarize other applications and issues in Wa/fC.

2.2.2 Grazing

In contrast to the safari model, Jeremy Whistle (1999) has his students graze in a pasture where he controls the kind and quality of the fodder. He has selected texts from the French Ministry of Foreign Affairs' online series "Label France". Intended for foreigners learning French, these texts are suitable in both language level and content, and obtaining permission from the ministry to incorporate them into an offline corpus for desktop use entailed no difficulties. Typically commercial sites require prior authorization for offline archiving and analysis. Since 1998 Knut Hofland has used his grazing permit from ten Norwegian newspapers to amass almost 400 million words of journalistic prose, identifying over a thousand "new" words (names, compounds and loanwords as well as neologisms) daily (http://avis.uib.no/). Similarly, GlossaNet (http://glossa.fltr.ucl.ac.be/) monitors 100 newspapers in 12 languages. Its publicly-accessible database is searchable by structure as well as word form, but unfortunately covers only several days' material.

With their explicit or implicit permission, official websites (e.g. http://www.un.org/documents/) and text archives (e.g. http://gutenberg.org) lend themselves to unrestricted grazing and archiving for offline use. For example, OPUS (Tiedemann and Nygaard 2004), an open source parallel corpus, collects, linguistically annotates and aligns parallel translated texts from the Web, primarily from freely available government sources. To extend the very productive focused grazing model from WfC to WaC, search agents like KWiCFinder can restrict searches to known sites with appropriate content and language to harvest texts for online concordancing or offline use.

3. Search Engines Past, Present and Future

3.1. Search Engines and Searchers

SEs remain key tools to find online documents to compile a corpus, and effective use of offline corpora requires search skills as well. Understanding how SEs work and how they are evolving to improve lay searchers' satisfaction is essential for serious

[9]Description of the project approach and the resulting *Wörterbuch Nationale Varianten des Deutschen* online at http://www.germa.unibas.ch/deusem/forsch/Prolex/prolex.de.html

exploitation of the Web as a corpus resource. Commerce drives today's Web, with significant consequences for online linguistic research. The large general-purpose search sites we must rely on are business ventures, developed and operated at enormous expense. They provide essential services in exchange for advertising fees, and "paid positioning" is intended to steer searchers away from more relevant "natural" search results toward advertisers' sites.

The average searcher's interests and requirements are quite different from those of a language scholar or learner. While the former wants to explore a question exhaustively, typical SE users have a specific content-oriented goal such as locating a specific site, finding valid information on a topic, or discovering a source for a Web-mediated product or service. In a classical paper drawing on his experience at AltaVista, Broder (2001) designates these goals as *navigational, informational* and *transactional* respectively. A user survey and analysis of actual queries at AltaVista (AV) identified the underlying information needs as 20% navigational, 48% informational and 30% transactional, with some overlap between the latter two categories.

Over the last decade, SEs have evolved away from demanding sophisticated searching skills from the user to boost results' relevance. What Broder calls first-generation Web search relied upon on-page information – search term salience in text frequency and formatting – and was best suited to full-text search for informational queries; epitomized by the string-matching power of AV, it represented state-of-the art through 1997. Second-generation SEs use off-page Web-specific information like PageRank, the link popularity ranking introduced by Google in 1998 as an indicator of page quality. By proving effective for both navigational and informational queries, this approach has made Google the market leader. Since the early 2000s, third-generation approaches have attempted to identify the "need behind the query" to identify relevant results – while providing targeted advertising. According to Broder, semantic analysis and context determination enable rapidly-evolving SE techniques to improve precision (relevance of search results) for all three kinds of queries.

3.2 Consequences of Current Trends in Web Search

Investigations of the typical user's preferences and search behavior have strongly influenced online searching.[10] Information seekers immediately confront the crucial problem of Information Retrieval (IR), maximizing both precision and recall, i.e. ideally matching *only* (precision) and *all* (recall) relevant documents. Two recent articles, Asadi and Jamali (2004) and Evans et al. (2005), sketch how SEs are evolving to address this problem. Continuing in the IR tradition, first-generation SEs supported sophisticated querying to boost result relevance. While AltaVista (AV) once imposed no limits on query length or complexity, complex queries were rare,

[10] Relevant user studies include Silverstein et al. (1999), Körber (2000), and Spink and Jansen's summary article and book-length synthesis (2004a, b). For a recent critical review of the literature see Martzoukou (2004). Bates (2002) contextualizes information seeking in humankind's behavior and evolution.

and up to 25% those submitted to AV were ill-formed and thus returned no results (Silverstein et al. 1999; Körber 2000). Currently, 80%-90% of all SE queries consist of a single word or very brief phrase, usually a noun, very frequently a proper noun, and in languages where this makes a difference, in the nominative form. Searches for other word classes are rare except in phrases.[11] The predominance of short, simple searches and improvements in result ranking schemes have permitted SEs to abandon underused "geek-seek" features with their high computational overhead such as nested bracketing, wildcards, long queries and large result datasets, and they have incorporated features like proximity, stemming and fuzzy matching into their standard matching algorithms. Unfortunately for language professionals, it is precisely such complex query tools that facilitate targeted online linguistic research.

The query, search and ranking optimization techniques SEs have adopted can either assist or sabotage a scholar's quest. On the positive side, when vague queries match large numbers of disparate documents, some SEs suggest query refinements based on frequently co-occurring terms,[12] which can improve the relevance of results upon re-query – and provides a ready list of typical collocates to boot. Geographic relevance is a ranking criterion with both pros and cons: SEs guess the user's location by IP address, then rank results (and display advertising) by presumed proximity to the searcher. While beneficial for marketing, this technique can interfere when investigating a foreign language. For example, when I seek English-language pages from a Dutch IP address, some SEs rank hits in the Netherlands higher than for the same search with an IP address in the U.S. via a VPN client. Automatic geographic ranking can undermine a quest for authoritative examples, but optional specification of the region to search would be useful. Finally, all major SEs now take link popularity into account to rank results. This sacrifices diversity in the search results, biasing them toward large, popular sites.[13]

3.3 The Future of Web Search

What will the next big developments in Web search bring? Major SEs will soon capitalize on document clustering and display techniques like those developed by Vivisimo, Kartoo, Ujiko and Grokker, which offer more meaningful ways to represent information and organize SERPs than a ranked listing of matching hits.[14] Labels

---

[11] http://searchenginewatch.com/facts/article.php/2156041 links to sites listing popular query terms; some display sample queries as they are being processed by SEs. http://wordtracker.com delivers a weekly list from various sources by e-mail, and similar services exist for other languages as well.

[12] For example, to help users focus queries, AlltheWeb offers lists of terms to select for inclusion or exclusion in a refined query. Similarly, when Google allows one to search for pages related to a given hit, and when it finds few hits for a query, it may propose a more frequent alternative with "Did you mean to search for _____?"

[13] Pandey et al. (2005) propose to "shuffle the stacked deck" by mixing a random sample into popularity-ranked results so less familiar sites gain exposure. For linguistic research the ability to tweak popularity weighting would be most useful.

[14] Document clustering discovers features shared by webpages and groups them together by those features; this is distinct from classifying documents by categories determined *a priori.* Additional clustering resources: Carrot2 (http://sourceforge.net/projects/carrot2/, demonstrator at http://carrot.cs.put.poznan.pl/), an open-source search results clustering framework; SnakeT, a

extracted automatically from the document clusters will provide linguists easily accessible, productive mines for lexical associations. Another SE trend is personalizing search by basing SERP ranking on analysis of patterns in the user's browsing and searching habits, an approach that could improve relevance for language-oriented (re)search. In addition, industry analysts expect significant growth in "vertical" search, i.e. specialized SEs dedicated to a single content domain or region, which will allow language professionals to target searches more precisely. Desktop search (DS) applications offered by major SEs are integrating offline and online search, eradicating distinctions between document locations and types of information resources. Thanks to their application programming interfaces (APIs), these technologies have tremendous potential as corpus tools.[15]

In the U.S., the major search sites have become the largest growth sector in the information economy, diverting advertising dollars away from print and other media.[16] For the typical searcher, SERPs from all the major search sites are now comparable in relevance and usefulness, so SEs must compete for market share on other grounds. They will continue to improve search functionality and add non-search features to their sites. Any successful enhancement will be copied by other major sites. In the future, user loyalty will derive more from inertia and dependence on other services (e.g. news, video, audio, e-mail, blogging, online photo albums, discussion group hosting) than from perceived search quality.

While SEs have little incentive to address language researchers' specific needs directly, the innovations and services introduced to boost competitiveness will benefit us ultimately. As they expand global coverage, SEs will spur development of natural language processing technology for a growing range of languages. Academic research across the spectrum of search and information science issues will expand, with opportunities for cross-disciplinary collaboration and funding—and employment for our graduates. Having mastered scaling databases to terabytes of data, SEs can now focus on discovering and relating patterns in those data, leading to new linguistic knowledge.

Efforts to build user loyalty by customizing the search experience are resulting in greater power and flexibility for those whose research rides on public SEs. Free APIs enable rapid incorporation of sophisticated search into special purpose application programs. Currently Yahoo offers the most varied API, supporting not only classical Web search, but also context search and term extraction from uploaded texts. One can even restrict results to content with a Creative Commons license, some of which could be reused for distributable corpora. Yahoo's My Web services even allow one to build, search and share online archives of webpages, an avenue to WfC requiring minimal technical sophistication for the user. Microsoft's search API allows one to tweak the settings for ranking and matching factors, reducing SE second-guessing which can degrade result quality for a linguist.

---

personal meta-search engine which clusters based on snippets from SERPs (Ferragina and Gulli 2005).

[16] Google and Yahoo's revenue grew from $2.5 billion in 2003 to $6.5 billion in 2004, and at this writing the total value of Google's stock is greater than that of any other "media" company.

4. Concordancing the Web

4.1 KWiCFinder Concordancing Search Agent

During the Web's early history SEs were disappointingly ineffectual. AV's launch in December 1995 changed that – and made me an intensive SE user. While I soon developed techniques and programs to maximize efficiency of downloading and evaluating webpages, few students or colleagues adopted my multitasking methods. To expedite finding and reviewing webpages, I programmed 16-bit KWiCFind, which excerpted documents and produced summary reports with keyword in context (KWiC) display, piloted in 1997. After a complete overhaul, 32-bit KWiCFinder (KF) premiered publicly at (Fletcher 1999). It has continued to evolve, and can be downloaded free from http://kwicfinder.com/.

4.1.1 KWiCFinder and AltaVista

When I developed KF, AV offered the most powerful full-text matching capabilities. Since AV was acquired and retired[17] by Yahoo! in the spring of 2004, much of that power was lost. We will review those features to highlight essential features for efficient search. AV-Y designates Yahoo's limited successor to AV.

AV indexed **all** words, even function and other high-frequency words ignored by some other SEs which may be the target of a linguistic investigation. AV-Y continues to indicate which webpages contain these "stopwords", but does not reliably track the co-text, reducing its usefulness for exact phrase matching. Formerly AV distinguished upper- from lower-case and "special" characters with diacritics from their "plain" counterparts, and incorporated language-specific knowledge, such as equivalence of *ä* and *ae, ß* and *ss* in German. Major SEs no longer support search by case, and support for query by special characters is either lacking (Google) or inconsistent on most SEs.[18]

Early advocates of WaC will remember AV for its innovation, power and size. It was the first to SE provide true world-wide multilingual coverage, and it introduced document clustering ("Live Topics"), search by language, webpage translation (Babelfish) and integrated desktop search. To support narrowly focused searches AV offered Boolean operators including NEAR (i.e. within 10 words of another search term), nested bracketing, and wildcards,[19] and imposed no limits on query length or complexity. It allowed matching any number of documents (current standard practice limits results to 1000 webpages), crucial for random sampling of the Web.

---

[17] While the site http://altavista.com still exists, it now uses the Yahoo! search database with greatly reduced full-text search capabilities. Ray et al. (1996) portray the exciting early days at AltaVista.

[18] http://forums.searchenginewatch.com/showthread.php?t=6013 explores special-character matching issues.

[19] Google does offer two kinds of wildcard matching: "wildwords", where * can match any word in a phrase, e.g. *the * * of* matches "the lower house of" etc.; the "synonym" operator ~, which matches alternate forms and semantically similar words, e.g. *~labor* matches "labour" as well, and *~nation* also matches "nations" and "national".

AV also set off the first SE size war by indexing 16 million pages at launch, while its competitors' databases boasted fewer than two million entries. Finally, AV enabled the first SE-based study in corpus linguistics, Brekke (2000).

Unfortunately AV's innovative search technology was "locked inside a dying company" which did not support it properly (Schwartz 2004). After surviving several changes of ownership and reorganization, AV-Y's market share has dropped well below 1%, and it might disappear entirely before long. Fortunately a noble successor to AV has appeared on the horizon. Exalead, a new Web SE based in France, supports all of AV's sophisticated features and much more, even offering regular expression pattern matching, with an API and destop search in the works. Once again the future appears bright for geek seek!

4.1.2 KWiCFinder's Enhancements to Web Search

For precisely focussed queries, KF offers matching strategies beyond AV's capabilities. AV(-Y) automatically matches a plain character in a search term with any corresponding accented character, and lower-case letters also match their upper-case counterparts (e.g. *a* in a search term matches any of *aáâäàãæåAÁÂÄÀÃÆÅ).* In typical Web searches these "implicit wildcards" ensure that paradigmatic and graphic variants of a given word match a single search term, despite factors like sentence-initial capitalization; required, omitted, or misused diacritics; or alternate spellings due to keyboard limitations.

Wildcards simplify *entering* search terms, but they also lead to irrelevant matches which must be eliminated individually. To address this problem I implemented single-character wildcards: *?* and *%,* which match either one (no more, no less) or zero to one character respectively. KF's *"sic"* option forces exact match to plain or lower-case characters in a query; without *sic,* the query *polish* matches *Polish* as well. Similarly, KF complemented AV's NEAR Boolean operator, with BEFORE and AFTER operators and specification of the distance between the terms.[20] Extensions like single-character wildcard and *sic* matching do come at a significant price: KF may have to retrieve, analyze and discard many documents matched by the SE which fail the user's finer criteria.

Completely specifying alternate forms makes searches more efficient than wildcard queries, but entering variants is time consuming. KF introduced "tamecards", a shorthand for alternate forms. For example, the tamecard query *s[iau]ng[,s,ing]* expands to all forms of the verb *sing: sing, sings, singing, sang, sung* (fortunately the nonsense forms *sangs, sungs, sanging, sunging* yield no false matches). The SE is queried for any of these forms, and only exact matches are processed. Since morphological patterns typically apply to many words, tamecards can be saved and pasted into queries as needed. A further refinement is the "indexed tamecard", in which every *n*th field in curly braces corresponds to the *n*th field in other sets of

[20] When AV still supported NEAR, it also had the undocumented ability to match pairs of terms within *any* specified range, and it supported BEFORE and AFTER as well, either with or without specifying proximity. (http://www.searchengineshowdown.com/features/av/index.shtml)

curly braces within the same search term, so that *{me,te,se} lav{o,as,a}* expands only to reflexive *me lavo, te lavas, se lava.*

Other KF tamecards address orthographic variants with or without hyphens or apostrophes. Search terms with this punctuation are expanded to alternate forms, so *on-line* matches any of the spellings *on-line, on line,* or *online,* and German *ich hab's* matches both *ich hab's* and *ich habs.* This shorthand is particularly useful for English, where national and individual usage varies, and German, now in transition to a new spelling. German reforms permanently separate many words once written as one, while fusing some former phrases into single words and permitting individual discretion in breaking up compounds. Thanks to KF's tamecards, queries like *kennen-lernen* match both old-style *kennenlernen* and reformist *kennen lernen* with a single entry.

Finally KF introduced "inclusion" and "exclusion" criteria, terms and conditions either to target a specific content domain or to disqualify documents from consideration. Such terms are added to the SE query to focus a search, but do not appear in KF's KWiC concordance report. Other selection criteria include date, Internet domain (a rough guide to country of origin), as well as host, i.e. a specific Web server, and URL.

4.1.3 KWiCFinder Concordance Reports

While processing a query, KF retrieves up to 20 documents a minute, excerpts them and produces a concordance of the key search terms in context, along with information about the source documents. The search reports are encoded in XML (eXtensible Markup Language) and offer a choice of interactive display formats through a set of XSLT (eXtensible Stylesheet Language Transformation) stylesheets. To display a useful report, KF transforms this XML with an XSLT "stylesheet" to select the information to show, insert text labels and format the result as an HTML document for browser display. To change the display layout or language, a different stylesheet is applied to the same XML data. With knowledge of XSLT and browser scripting techniques, an end user can create new report formats or apply other stylesheets to annotate, merge, prune, or restructure XML search reports.

JavaScript and dynamic HTML enable substantial interactivity in KF's classical KWiC display reports. The user can specify criteria for re-sorting concordances and tallying forms in the co-text. Searching for specific forms tallies and highlights them, and buttons provide rapid navigation between highlighted forms. Concordance lines can be annotated or deleted. Any user modifications to the KF report can be saved as a browser-based stand-alone interactive concordance. This approach points the way to a light-weight cross-platform solution for learner concordancing.

4.1.4 KWiCFinder and Web as / for Corpus

While the Web is no corpus in the classical sense, I regularly access it with KF to research linguistic questions and to develop language instruction materials. Fletcher

(2004a) illustrates how with many concrete examples. Concordancing techniques are also beneficial at the text level, to evaluate content and form of webpages matching a query.

KF can also compile an ad-hoc corpus from the Web for offline analysis and use. For example, to build a sample Web corpus I recently ran up to 20 independent KF searches simultaneously on a home broadband connection. In one morning KF downloaded over 22 thousand webpages (9 GB of HTML) totaling 115 million words and saved them on my hard drive in text format. With techniques and software described elsewhere (Fletcher 2004b), I eliminated "uninteresting" and duplicate documents, leaving a 38 million word corpus for further processing into a database of *n*-grams and phrase-frames with full-text KWiC concordances available on demand.[21] While this project took almost a day, a corpus of a few million words can be completed in less than an hour.

4.2 WebKWiC

For searchers who prefer not to install KF I developed WebKWiC (WK),[22] a fully browser-based JavaScript application. It takes advantage of Google's "Document from Cache" feature to automate webpage retrieval and markup: Google serves up copies of webpages from its archives, highlighting the search terms with color codes. WK retrieves these cached pages in batches, adding buttons for easy navigation among highlighted terms and windows. WK provides an interface for special character input and gives essential search options greater prominence than does Google's original page. Google is an ideal partner for an entry-level search agent like WK. Its straightforward approach to advanced search with "implicit Booleans" is easy to learn and widely imitated, so users either come with or acquire readily transferable skills. Since Google indexes major non-Western European / non-Latin orthography languages, WK meets needs which KF does not address.[23]

5. Webidence as Linguistic Evidence

We all know the limitations of online information: there is too much ephemeral content of dubious reliability; journalistic, commercial and personal texts of unknown authorship and authority abound; assertions are represented as established fact, and details of sources and research methodology are documented haphazardly at best. For linguistic research even more caution is essential. The Internet domains in a URL (.ca, .uk, .de, .jp, .com, .net etc.) are at best a rough guide to provenance. Furthermore, many webpages are mainly fragments–titles and captions, with the occasional imperative ("click here", "buy now"). As the lingua franca of the digital frontier, English is both the target and source of contamination: non-Anglophones often translate their webpages into Info-Age pidgin English while fusing creolized Web English into texts in their native tongue. Similarly, searches for linguistic

[21] kfNgram and companion database builder / browser software, also free from KWiCFinder.com.
[22] http://kwicfinder.com/WebKWiC/
[23] Fletcher (2004a) compares and discusses alternatives to KWiCFinder in depth.

examples can lead to work by learners with imperfect mastery of the language or to baffling machine translations. In many online forums, careless or cryptic language and sloppy spelling prevail. With its frenetic pace of development, the Web typically values content *creation* above *perfection* and tolerates ill-formed language–anyone upset by this is but a click away from relief.

5.1. Search Engines as Gateways to Webidence

In light of these pitfalls we need "Standards of Webidence" to guide the selection and documentation of online language for linguistic research. We also must understand and beware of SEs' limitations. In particular, hit counts reported by a SE give only a general indication; these numbers cannot prove the prevalence or appropriateness of a given formulation.[24] SEs warn not to trust their figures, and with good reason: generating SERPs receives priority over estimating hit counts, and the exact form and order of search terms affects those counts. For several reasons numbers for the same or equivalent query easily vary up to an order of magnitude.[25] Moreover, SEs report *document count,* i.e. **the number of webpages** matching a query, **not the actual number of occurrences** on those pages. A single document may contain alternate usages, thus appearing in multiple counts, and numerous pages propagate verbatim a formulation originating in a single document, thus multiplying its apparent frequency. Some spurious or unusual usages are traceable to a single source. Fletcher (2004b) evaluates several approaches to filtering out "noise" resulting from highly repetitive, virtually identical and primarily fragmentary documents.

SE indexing and ranking practices also affect the usefulness of Web data. For example, Google proudly states "Searching 8,058,044,651 web pages", but it does not index all the text on so many pages. For a sizeable subset, analysis is limited to the hyperlink text and target. Moreover, following standard practice, only the first 100,000 words are indexed on any page. Since major SEs provide access to only the first thousand hits, the order of search results is crucial. Exact ranking criteria and weighting are continually tweaked proprietary secrets, but prevalent practice relies heavily on *link popularity* – the number and reliability of links to a webpage, indicative of authority and quality– and *term salience* expressed as *TF/IDF,* a metric derived from the ratio of the frequency of a search term (TF) in a given document to the inverse of the total number of documents in which it occurs (IDF). Yahoo and MSN apparently assign relatively more weight to term salience and less to link popularity than Google. For some purposes, however, a linguist requires texts in which a term simply occurs without being salient, as in this example. Recently a Dutch colleague asked what abilities one can *hone* metaphorically in English besides *skills.* The BNC offers only a handful of examples, so I went to the Web. The first

[24] It is unclear how many linguists understand these limitations. Postings in scholarly forums like Corpora List and Linguist List citing evidence from SE hit counts rarely indicate whether the poster has verified a substantial number of the hits or is even aware of the limitations of this method.
[25] Véronis (2005) documents striking discrepancies in Google hit counts for equivalent queries. Numerous threads on http://forums.searchenginewatch.com/ detail factors in such variation.

thousand hits reminded me that in the commercial world hones collate with knives and chisels, brakes and engines, stone and tile, but few abilities appeared. Repetition made variants of *hone* salient on pages offering such products and services; I had exhausted my quota with few metaphoric results. In contrast, my randomly compiled Web corpus (4.1.4) has more relevant examples, and almost none of the concrete use.

5.2 Verifiability: Preserving and Sharing Webidence

Verifiability is a cornerstone of responsible research: evidence for any claim or conclusion must be subject to inspection and alternate analysis by other researchers. The Web's volatility diminishes its credibility for research. Not only do hit counts vary widely due to non-linguistic factors, but the same query on the same search site can return different sets of SERPs, not only from different places or at different times, but even during a single user session. In the best case, the laws of large numbers permit *comparable* results for frequent search terms, but the composition of the actual webpages matched can be quite different.

No Web search data are truly verifiable by other investigators, one reason I propose a Web Corpus Archive (Fletcher 2004a). A few principles would represent progress toward that goal for WaC research. Investigators should make all webidence accessible to others for verification or reuse, preferably online. If SE hit counts are used, multiple SEs should be queried with various search term orders, and the queries should be rerun at several week intervals and on different regional versions of the SE to ensure stable counts and tolerable variance; the corresponding SERPs would be retained as webidence. Webpages on which an analysis is based must be preserved and shared, as should other matching pages.

KF facilitates responsible online linguistic scholarship in several ways. One can review large numbers of documents and concordances efficiently. Each keyword is displayed in sufficient context to evaluate its relevance and validity, and the total number of occurrences can be tallied. Webpages can be saved locally for further analysis or independent verification of results. Complementary corpus tools can process these webpages to eliminate repetitive or redundant documents, to analyze lexical patterns, and to compile databases for further exploration and deployment on the Web.

6. The Future of the Web for / as Corpus

Recent developments inspire considerable optimism about the prospects for Wa/fC. Major SEs are introducing services and features that lower the threshold for simple Web concordancing and archiving for e.g. translators and language teachers. New SEs even improve on the level of search sophistication we once enjoyed with AV. Thanks to powerful free tools for customizing every aspect of crawling, analyzing, searching and archiving Web documents, Wa/fC linguists can focus on their research, not on studying Internet protocols and developing software from scratch. At least two research groups—the University of Central England's RDUES WebCorp

and the WaCky consortium organized by the University of Bologna-Forlì's SSMILT—are working toward multi-language SEs for linguists, and other Wa/fC projects are underway. The interests of corpus and computational linguists are intersecting in novel ways with those of computer and information scientists, suggesting broader opportunities for fruitful collaboration and funding. As practices evolve to ensure the integrity of Web data, it will become fully accepted as a legitimate source for linguistic research. This explosion of activity in Wa/fC a decade after the Web's Big Bang promises ongoing innovation and ample rewards as we apply this boundless resource to our endeavors.

REFERENCES

Links verified May 2005; * means no longer at the URL given, but still available on the Wayback Machine http://web.archive.org.

Asadi, Saeid and Hamid R. Jamali. (2004). Shifts in search engine development: A review of past, present and future trends in research on search engines. *Webology,* 1(2), Article 6.
http://www.webology.ir/2004/v1n2/a6.html

Banko, Michele and Eric Brill. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. ACL 2001.
http://research.microsoft.com/~brill/Pubs/ACL2001.pdf

Bates, Marcia J. (2002). Toward an Integrated Model of Information Seeking and Searching. Fourth International Conference on Information Needs, Seeking and Use in Different Contexts, Lisbon, Portugal, September 11-13, 2002.
http://www.gseis.ucla.edu/faculty/bates/articles/info_SeekSearch-i-030329.html

Bickel, Hans (2000). Das Internet als Quelle für die Variationslinguistik. In: Annelies Häcki Buhofer (ed.), Vom Umgang mit sprachlicher Variation. Soziolinguistik, Dialektologie, Methoden und Wissenschaftsgeschichte. Festschrift zum 60. Geburtstag von Heinrich Löffler, 111-124. Tübingen: Francke.
http://www.germa.unibas.ch/seminar/whoiswho/Publikationen/Variationsling.pdf

Bickel, Hans and Regula Schmidlin. (2004). Ein Wörterbuch der nationalen und regionalen Varianten der deutschen Standardsprache. In: Thomas Studer and Günther Schneider (eds.), *Deutsch als Fremdsprache und Deutsch als Zweitsprache in der Schweiz, Bulletin valsasla* 75, 99-122.
http://www.germa.unibas.ch/seminar/whoiswho/Publikationen/BickelSchmidlin.pdf

Brekke, Magnar. (2000). From the BNC toward the Cybercorpus: A Quantum Leap into Chaos? In Kirk, John. M. (Ed.), Corpora Galore: Analyses and Techniques in Describing English. Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora (ICAME 1998), 227-247 Amsterdam, Atlanta: Rodopi.

Broder, Andrei. (2002). A taxonomy of web search. *ACM SIGIR Forum Archive* 36(2), 3-10.
http://www.acm.org/sigir/forum/F2002/broder.pdf

Carlson, Lauri, Minna Grönroos and Suvi Lemmilä. (2005). Squirrel Two: Experiments on a metasearch engine for CALL. NODALIDA 15, Joensuu, Finland, 20-21 May 2005.
http://phon.joensuu.fi/nodalida/abstracts/28.shtml

Chakrabarti, Soumen. (2002). Mining the Web: Analysis of Hypertext and Semi-Structured Data. San Francisco: Morgan Kaufmann.

Chesñevar, Carlos I. and Ana G. Maguitman. (2004). An Argumentative Approach to Assessing Natural Language Usage based on the Web Corpus. *Proceedings of the European Conference on Artificial Intelligence (ECAI).* Valencia, Spain, 22-27 August 2004.
http://fermat.eps.udl.es/~cic/2004/2004_ecai.pdf

De Schryver, Gilles-Maurice. (2002). Web for / as corpus: a perspective for the African languages. *Nordic Journal of African Studies* 11(2), 266-282.
http://tshwanedje.com/publications/webtocorpus.pdf

Dura, Elżbieta. 2004. Concordances of Snippets. Workshop "Enhancing and using electronic dictionaries", COLING, Geneva, August 2004.
http://82.182.103.45/Lexware/English/publications/coling04.pdf

Evans, Michael P., Richard Newman, Timothy Putnam and Diana J.M. Griffiths. (2005). Search Adaptations and the Challenges of the Web. *IEEE Internet Computing* 9(3), 19-26.
http://doi.ieeecomputersociety.org/10.1109/MIC.2005.65

Ferragina, Paolo and Antonio Gulli. (2005). A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering. *Proceedings of WWW 2005*, 10-14 May 2005, Chiba, Japan, 801-810.
http://www2005.sfc.keio.ac.jp/cdrom/docs/p801.pdf

Fletcher, William H. (1999). Winnowing the Web with KWiCFinder. CALICO, Miami University of Ohio, Oxford, OH, 5-9 June 1999.

Fletcher, William H. (2001). Concordancing the Web with KWiCFinder. American Association for Applied Corpus Linguistics, Third North American Symposium on Corpus Linguistics and Language Teaching, Boston, MA, 23-25 March 2001.
http://kwicfinder.com/FletcherCLLT2001.pdf

Fletcher, William H. (2004a). Facilitating the compilation and dissemination of ad-hoc Web corpora. In Guy Aston, Silvia Bernardini and Dominic Stewart (eds.), *Corpora and Language Learners*, 271-300. Amsterdam: John Benjamins. Series: Studies in Corpus Linguistics 17.
http://kwicfinder.com/FletcherTaLC5.pdf

Fletcher, William H. (2004b). Making the Web more useful as a source for linguistic corpora. In Ulla Connor and Thomas A. Upton (Eds.) *Applied Corpus Linguistics. A Multidimensional Perspective,* 191-206. Amsterdam/New York: Rodopi. Series: Language and Computers - Studies in Practical Linguistics 52.
http://kwicfinder.com/AAACL2002whf.pdf

Fletcher, William H. (2005). Towards an Independent Search Engine for Linguists: Issues and Solutions. Symposium "La Rete come Corpus", University of Bologna, Forlì, Italy, 14 January 2005.
http://kwicfinder.com/WaCForli2005-01.pdf

FUNREDES. (2001). The Fifth Study on Languages and the Internet.
http://funredes.org/LC/english/L5/L5contents.html

Gerbig, Andrea and Patricia Buchtmann. (2003). Vom "Waldsterben" zu "Geiz ist Geil": Figurativer Sprachgebrauch im Paradigmenwechsel von der ökologischen zur ökonomischen Handlungsmotivation. *metaphorik.de* 04, 97-114.
http://www.metaphorik.de/04/gerbigbuchtmann.pdf

Global Reach. (2004). Global Internet Statistics (by Language).
http://global-reach.biz/globstats/

Grefenstette, Gregory and Julien Nioche. (2000). Estimation of English and non-English Language Use on the WWW. RIAO 2000, Paris, 12-14 April 2000.
http://arxiv.org/ftp/cs/papers/0006/0006032.pdf

Gulli, Antonio and Alessio Signorini. (2005). The Indexable Web is More than 11.5 Billion Pages. *WWW 2005,* May 10–14, 2005, Chiba, Japan.
http://www2005.sfc.keio.ac.jp/cdrom/docs/p902.pdf

Hawkins, Donald T. (1996). Hunting, Grazing, Browsing: A Model for Online Information Retrieval. *ONLINE* 20(1).
http://web.archive.org/web//http://www.onlinemag.net/JanOL/hawkins.html

Ho, James. (2005). Hyperlink obsolescence in scholarly online journals. *Journal of Computer-Mediated Communication,* 10(3), article 15.
http://jcmc.indiana.edu/vol10/issue3/ho.html

Keller, Frank and Mirella Lapata. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29(3), 459-484.
http://acl.ldc.upenn.edu/J/J03/J03-3005.pdf

Kilgarriff, Adam and Gregory Grefenstette. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3), 333-347.
http://acl.ldc.upenn.edu/J/J03/J03-3001.pdf

Körber, Sven. (2000). Suchmuster erfahrener und unerfahrener Suchmaschinennutzer im deutschsprachigen World Wide Web. Ein Experiment. Unpublished master's thesis, Westfälische Wilhelms-Universität Münster, Germany.
*http://kommunix.uni-muenster.de/IfK/examen/koerber/suchmuster.pdf

Lawrence, Steve and C. Lee Giles. (1999). Accessibility of Information on the Web. *Nature, 400:* 107-109. Summary, commentary, update and download at *http://web.archive.org/web/*/http://www.wwwmetrics.com

Martzoukou, Konstantina. (2004). A review of Web information seeking research: considerations of method and foci of interest. *Information Research,* 10(2), paper 215.
http://InformationR.net/ir/10-2/paper215.html

Mikami, Yoshiki and Izumi Suzuki. (2005). The Language Observatory Project and its Experiment: Cyber Census Survey. Crossing the Digital Divide. SCALLA 2004.
http://www.elda.org/en/proj/scalla/SCALLA2004/mikami.pdf

Mikami, Yoshiki, et al. (2005). The Language Observatory Project (LOP). *WWW 2005,* May 10–14, 2005, Chiba, Japan.
http://www2005.sfc.keio.ac.jp/cdrom/docs/p990.pdf

Moore, Alvin and Brian H. Murray. (2000). Sizing the Internet. 10 July 2000. Arlington, VA: Cyveillance, Inc.
*http://www.cyveillance.com/resources/7921S_Sizing_the_Internet.pdf

Nilsson, Kristina and Lars Borin. (2002). Living off the land: The Web as a source of practice texts for learners of less prevalent languages. *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation,* Las Palmas: ELRA, 411-418.
http://web.archive.org/web//http://fenix.ling.uu.se/lars/pblctns/lrec2002.pdf

Ntoulas, Alexandros, Junghoo Cho and Christopher Olston. (2004). What's New on the Web? The Evolution of the Web from a Search Engine Perspective. *WWW 2004*, ACM Press, 1–12.
http://www2004.org/proceedings/docs/1p1.pdf

Ntoulas, Alexandros, Petros Zerfos and Junghoo Cho. (2005). Downloading Textual Hidden Web Content by Keyword Queries. In *Proceedings of the Joint Conference on Digital Libraries* (JCDL), June 2005.
http://oak.cs.ucla.edu/~cho/papers/ntoulas-hidden.pdf

Pandey, Sandeep, Sourashis Roy, Christopher Olston, Junghoo Cho, Soumen Chakrabarti. (2005). Shuffling a Stacked Deck: The Case for Partially Randomized Ranking of Search Engine Results. In *Proceedings of 31st International Conference on Very Large Databases (VLDB),* September 2005.
http://oak.cs.ucla.edu/~cho/papers/cho-shuffle.pdf

O'Neill, Edward T., Brian F. Lavoie and Rick Bennett. (2003). Trends in the Evolution of the Public Web 1998 – 2002. *D-Lib Magazine* 9 / 4 (April).
http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

Ray, Eric J., Deborah S. Ray and Richard Seltzer. (1996). *The AltaVista Search Revolution.* Berkeley, CA:Osborne-McGraw-Hill.

Robb, Thomas. (2003). Google as a Quick 'n' Dirty Corpus Tool. *TESL-EJ,* 7(2), On the Internet.
http://www-writing.berkeley.edu/TESL-EJ/ej26/int.html

Salzmann, Anne and Douglas Mills. (1995-2005). LinguaCenter Grammar Safari.
http://www.iei.uiuc.edu/student_grammarsafari.html

Schwartz, Barry. (2004). "Search Memories – Live from SES San Jose". Search EngineWatch "SEM Related Organizations & Events", 8 May 2004.
http://forums.searchenginewatch.com/showthread.php?t=949

Silverstein, Craig, Monika Henzinger, Hannes Marais and Michael Moricz. (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33(1), 6 -12.
http://www.acm.org/sigir/forum/F99/Silverstein.pdf

Spink, Amanada and Bernard J. Jansen. (2004a). A study of Web search trends. *Webology,* 1(2), Article 4.
http://www.webology.ir/2004/v1n2/a4.html

Spink, Amanda and Bernard J. Jansen. (2004b). *Web Search: Public Searching of the Web.* Dordrecht: Kluwer, Information Science & Knowledge Management 6.

Tanguy, Ludovic and Nabil Hathout. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir duWeb. TALN 2002, Nancy, 24–27 June 2002.
http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/TALN/AC_0072.txt.html

Tiedemann, Jörg and Lars Nygaard. (2004). The OPUS corpus - parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal, May 26-28.
http://stp.ling.uu.se/~joerg/paper/opus_lrec04.pdf

Véronis, Jean. (2005). Web: Google's missing pages: mystery solved? Blog entry 8 February 2005.
http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html

Way, Andy and Nano Gough. (2003). wEBMT: Developing and Validating an EBMT System using the World Wide Web. *Computational Linguistics* 29(3), 421-457.
http://www.computing.dcu.ie/~away/PUBS/2003/way02.pdf

Whistle, Jeremy. (1999). Concordancing with Students Using an "Off-theWeb" Corpus. *ReCALL 11*(2), 74-80.

http://www.eurocall-languages.org/recall/pdf/rvol11no2.pdf

Yamanoue, Takashi, Toshiro Minami, Ian Ruxton and Wataru Sakurai. (2004). Learning Usage of English KWICly with WebLEAP/DSR. *Proceedings of the 2nd International Conference on Information Technology for Application (ICITA 2004)* http://attend.it.uts.edu.au/icita05/CDROM-ICITA04/papers/14-6.pdf