Concordancing the Web with KWiCFinder

William H. Fletcher
United States Naval Academy

American Association for Applied Corpus Linguistics
Third North American Symposium on
Corpus Linguistics and Language Teaching, Boston, MA,
23-25 March 2001

Revised Draft[1]
Please Do Not Quote Without Permission!!

fletcher@kwicfinder.com

Size and Composition of the Web

The World Wide Web is a wondrous place, with an overwhelming variety of information in countless languages and domains. Just how many webpages there are and how they are distributed by language and content are not easy questions to answer. The Web is constantly growing and changing, 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 extrapolates from different samples of an 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-2001 falls in the range of three to five billion (i.e. thousand million = 10^9), a number projected to grow to 10-15 billion by mid-decade; others believe these figures too conservative and place the actual numbers at two to three times as many.

These two billion-plus pages are only the visible tip of the iceberg. For a page to be indexable, there must be a valid HTML link to it from another publicly accessible site, which excludes the many pages with restricted access. Far larger is the vast "invisible web" of content in databases, which can only be evoked by entering relevant queries in a text box, and text materials stored in formats which are not typically indexed, such as word processor, Post Script and Adobe Acrobat files.[2]

Despite the overall size of this corpus, one language, English, continues to predominate. Studies conducted in 2000 by Inktomi and Cyveillance conclude that over 85% of publicly-accessible webpages are in English, but here again even the best-informed estimates vary widely. In the summer of 2001 the Agence de la Francophonie released *L5: the Fifth Study of Language and the Internet,* based on these studies and the one by Global Reach cited below, complemented by research into the numbers of webpages in various languages returned by search engines. This report investigates the relative presence of the Romance languages, German, and English among online documents. It shows strong growth among the non-English languages in the proportion of webpages found relative to English, concluding that the number of webpages in each is roughly proportional to the number of Web users with that language as native tongue. Data from these and other studies of linguistic diversity on the Web are

---

[1]This rough draft was originally submitted for publication in a volume of selected papers from the conference, a project apparently has been abandoned. It will be updated and revised, then submitted for publication elsewhere.
[2]Google.com has recently started indexing other online document formats such as Adobe Acrobat, PostScript, Microsoft Word and PowerPoint, and continues to add more word processor and spreadsheet file types; unfortunately, incomplete stripping of formatting codes and imperfect reconstruction of text in columns can interfere with word matching and automatic text analysis.

summarized in this note.[3]

Historically English-language users and content have overshadowed other languages on the Internet, but the trend away from the preponderance of English seems clear. Statistics compiled by Global Reach illustrate the long-term development. In 1996, four-fifths of the 50 million Internet users were native speakers of English. By September 2001 Anglophones constituted only 43% of the world's online population of 503 million. Global Reach expects their share to fall below 30% of the 850 million Web users projected for 2005.[4] The anticipated phenomenal growth in this non-Anglophone Web population should spur tremendous expansion of online resources in tongues other than English, particularly the smaller non-Western ones, to the benefit of those who teach, learn, and investigate these languages.

The Web as a Corpus for Language Learning

The abundant and varied texts of the World Wide Web tantalize linguists and language instructors alike: the Web's ever-expanding, self-renewing machine-readable body of Web pages in scores of languages are easy to retrieve, but they are also challenging to sift through and exploit efficiently. Yet there are compelling reasons to supplement existing corpora with online materials. Once compiled, a corpus represents a snapshot of language usage and issues at the time the content was produced. The great expense of setting up a large corpus precludes frequent replacement or updating, and content can age surprisingly quickly. In contrast, countless new documents appear on the Web daily, so examples of current language usage and contemporary issues abound. In addition, even a large corpus might include few examples if any of a relatively infrequent expression or construction that would not be difficult to locate online. Furthermore, certain domains or text genres may be underrepresented or missing entirely in an existing corpus. Using the Web as a source one easily can compile an ad-hoc corpus to meet the specific needs of groups of learners or translators. Finally, while off-the-shelf corpora and corpus tools may entail significant fees and often require expensive hardware, the Web is virtually free, and desktop computers to perform the necessary processing are now within the reach of researchers and students alike.

While the Web does not constitute a corpus in the classical sense, as a foreign language instructor I use

---

[3] **Percentage of webpages by language.** Based on Google.com's current figures, Alex Franz (2001) reports the following distribution of webpages, round to the nearester full percent: English 58%, Japanese 9%, German 7%, Chinese 6%, French 4%, Russian 3%, Spanish 3%, Italian 2%, Korean 2%, Portuguese 2%, other 4%. The Cyveillance study based on a sampling of 350 million webpages estimates that as of 10 July 2000, 84.7% of webpages were U.S.-based and the rest "international"; presumably many of the remainer would be in English as well (Moore and Murray 2000). In a telephone interview on 8 November 2000 Julie Keslick of Inktomi indicated that language count was not a primary goal of the January 2000 Web Map study. According to the Inktomi tally, the top ten languages were: English 86.55%, German 5.83%, French 2.36%, Italian 1.55%, Spanish 1.23%, Portuguese 0.85%, Dutch 0.54%, Finnish 0.50%, Swedish 0.36%, Japanese 0.34%. Since the figures add up to about 100%, these languages apparently were the only ones identified. Grefenstette and Nioche (2000) offer a methodologically interesting study to estimate the number of *words* (not webpages) in a number of Latin-alphabet European languages, but it makes no attempt to estimate numbers for other languages.

[4] **Number of Internet users by language.** Global Reach frequently updates its estimates of the global online population and fully discloses the methodology used to derive them. Its data from September 2001 show the following percentages of users for the top ten languages: English 43%, Chinese 9.3%, Japanese 9.2%, Spanish 6.7%, German 6.7%, Korean 4.4%, Italian 3.8%, French 3.3%, Portuguese 2.5%, Dutch 2.2%, Other 8.9%. Another interesting source of this data is the Nua Internet "How Many Online" page listed in the bibliography. Current estimates of number of webservers and users per country (not identical to the number per *language)* can be found at http://www.netsizer.com/daily/TopCountry.html.

my concordancing Web search agent KWiCFinder (described in detail below) several times a week to access it as if it were one, for developing instructional materials and as well as for my own edification and research, at both the micro- and macro-levels. Let me illustrate how with a number of concrete examples.

To teach grammar or vocabulary, the Web is my primary source of eloquent examples. When I examine examples from the Web, they often force me to refine my understanding of how the language works. For example, textbook chapters on German conjunctions always teach the contrast between *aber* 'but, however' and *sondern* 'but, rather'; I have tired of the small range of examples of *sondern* I can contrive for my students. When I turned to the Web for variety, a sample of 50 passages with *sondern* included none of the pattern "not A but B". In most cases *but* or even *rather* were not acceptable translations for *sondern;* instead, a reformulation with *instead* (as in this sentence) was most appropriate. (Once again one wonders why textbooks concentrate on the *least* frequent use and ignore the others?)

The Web also allows me to verify current and possible usages and to obtain a rough indication of their relative prevalence and distribution. For example, I was astonished to encounter the phrase *los sesentas* 'the sixties' in a Latin-American text; I teach and normally would expect *los (años) sesenta,* without the plural marker on the numeral; the former has the overtones of an Anglicism. A series of KWiCFinder searches revealed that this usage is common and locally predominant in Latin America, yet virtually unattested in Spain.

The Web also permits my students and me to confirm and acquire vocabulary not yet found in dictionaries – nor in off-the-shelf corpora. Once I came across the word *privacidad* 'privacy' in Spanish-language software instructions. My unabridged dictionary from the early nineties was ignorant of this neologism, so I again suspected a blatant Anglicism, but a KWiCFinder search proved that it is indeed used throughout the Spanish-speaking world, even by authoritative sites like IBM and Microsoft. In another case, when I was invited to give a keynote speech on technology in Dutch at a conference in Belgium, I initially felt insecure: while I have near-native fluency in the language, I have not kept up with the vocabulary of technology. By reading excerpts from webpages that dealt with related topics I could fill in the lexical gaps with minimal effort.

Working with an ad-hoc corpus can help students develop discovery skills and reinforce linguistic content. For instance, one intermediate German textbook I have used taught the passive voice – formed most frequently with the auxiliary *wurde* + past participle – right after the subjunctive, which is formed with *würde* plus infinitive for most verbs. Anglo learners tend to disregard both diacritics and details of form, so some students became confused. To help contrast the two constructions I built a pair of keyword-in-context (KWiC) concordances on *wurde* and *würde* plus personal endings from the Web and had the students analyze these constructions in context.[5] This enabled them to understand and contrast the building blocks of these two constructions better and to observe dozens of examples of each in action.

Concordancing techniques are also beneficial at the text level. When searching for online documents which will be linguistically accessible to my students, I display the query terms in large chunks of context, up to a couple of hundred words. These long excerpts enable me to evaluate the language and content of the texts quite efficiently. I have learned to skim through excerpts from scores of documents and identify the most appropriate ones quickly. This text-level approach is useful for discerning both content and form, i.e. documents on a given topic or from a desired domain as well as those exemplifying

[5]KWiCFinder supports exporting search reports to an HTML file. Various interactive tools programmed in JavaScript are incorporated into this file, permitting browser-based stand-alone analysis; for details refer to http://miniappolis.com/KWiCFinder/KWiCFinderKWiCFeatures.html.

a given construction or register. With judicious choice of search terms one can locate texts rich in e.g. past tense forms or subjunctives to serve as readings to reinforce acquisition of structures and forms. Students can also follow this technique to locate relevant online resources for Web-based reports and research projects. Those who do tend to consult a greater variety of sources; those who choose not to often rely on the first few links found on Yahoo or Google.

Approaches to Exploiting the Web as a Corpus

A well-known model for finding and using information distinguishes three basic approaches: *hunting,* or searching directly for specific information, *grazing,* or using ready-made data sets which are composed and updated by an information provider, and *browsing,* or finding useful information by serendipity.(Hawkins 1996) Each of these approaches can serve as a model for corpus building or utilization; a melding of these techniques is most typical-- and most successful.

HUNTING

Due to the Web's size and lack of organization a search engine provides the most effective entry point for hunting information. There are dozens of general Search engines with world-wide reach, and thousands of others which concentrate on specific geographic regions, knowledge domains, or languages. Since the dawn of Web civilization, Anne Salzman and Doug Mills have sent their ESL students at the University of Illinois on "Grammar Safaris". With their online assignment sheets as guide and armed only with a web browser, they use a Search engine to track down webpages with examples of the structures they are studying. Then they use the browser's Find function to locate the examples within the documents and they copy and paste them into a word processor document to bring them to class. According to Salzman and Mills (2001), this approach from the Info-Stone-Age yields plenty of meat for classroom discussion. The hunting model is also being followed to exploit the Web as a corpus for linguistic research. Hans Bickel (2000) reports that investigators at the universities of Basel, Duisburg, and Innsbruck are trolling the Web for examples of regional usage in the various German-speaking countries to complement the material they have gleaned from other sources.[6] Other powerful solutions built 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), and, building on Grefenstette's proposed techniques, to resolve prepositional phrase attachment ambiguities for parsing (Volk 2000, 2001).

GRAZING

A hunting party sometimes returns empty-handed, and how much time and effort it will take to bag useful citations is rarely predictable. In contrast to the safari model, Jeremy Whistle (1999) turns his students loose in a ready-to-graze pasture where he controls the kind and quality of the fodder that awaits them. He has selected texts from the "Label France" series published online by the French Ministry of Foreign Affairs. Since these texts are intended for foreigners learning about French civilization and culture, both the language and content is suitable for his students. As government-sponsored instruments of cultural diffusion, the documents entailed no difficulties in obtaining the rights to incorporate them into an offline corpus for desktop use. (The question of developing offline corpora from online documents is addressed extensively below.) With a search agent like KWiCFinder this approach could easily be implemented in

[6]See the project description "Wörterbuch Nationale Varianten des Deutschen" online at http://www.germa.unibas.ch/deusem/forsch/Prolex/prolex.de.html

the online-mode: searches could be restricted to a known site or range of sites with appropriate content and language. This extends the very focused and productive grazing model to webpages for which one cannot (or lacks the time to) obtain permission for offline use.

BROWSING

Browsing is central to the Web – indeed, the unplanned discovery of information and insights lies at the heart of learning and research, and both hunting and grazing demand fortuitous finds to succeed. When consciously searching the "World Wide Haystack", most experienced "hunters" use search-engine hits merely as a point of departure for further browsing; they then typically follow several additional layers of links before reaching their goal. (Körber 2000) My hard drive preserves scores of documents I have chanced upon while looking for something else online and have saved for possible use in teaching or research. More frequently I rely on the "applied serendipity" approach described above: sending a search agent to retrieve and excerpt large numbers of documents, then scanning the results to winnow out the chaff and keep the grain. Silvia Bernardini (2001) has written of a systematic approach to increasing the number of serendipitous finds by having students work with a number of different corpora and analytical tools. Jennifer Pearson (2000) stresses that one must guide students to recognize true serendipity, i.e. to determine consciously whether an online document meets the essential criteria of reliability and appropriateness for one's purposes, in this case to serve as a model for translation[7]

Search Engines Present and Future

Unless one has already chanced upon suitable pastures for grazing, Search engines remain an essential tool for building any extensive corpus of online documents. The challenges are to ensure that a search yields maximally relevant results and to separate out irrelevant and uninteresting documents efficiently.

The dynamic, increasingly market-driven nature of the Web entails significant challenges and frustrations for efficient online concordancing. The large general-purpose search sites are commercial ventures, set up and maintained at enormous expense. They exist to generate advertising and sales revenues for their owners in exchange for providing a useful service. Merely by coincidence they also can serve serious research purposes, but their owners have no incentive to address the specific needs of academics. In order to maintain or attain profitability, many search sites are evolving into marketing sites: through policies of paid inclusion or paid positioning they can steer searchers away from more relevant results toward their advertisers.

Search engines target the average searcher, whose requirements are quite different from those of a scholar or student. Casual users typically have a well-defined information need such as locating a specific site, finding a valid answer to a question, or finding a well-stocked site meeting their search criteria. In contrast, scholars and teachers must examine and evaluate a range of resources to find the most reliable sources and the most useful texts. Search engines excel at returning *large numbers* of hits (documents matching one's query), but not at optimizing their *relevance* to the searcher's intent. Frequent changes in document content and "link rot"—the tendency of webpages to move without a forwarding address, or disappear from the Web altogether—can diminish the usefulness of search results even further.[8]

---

[7]In a very revealing study Hildreth (2001) investigated the factors underlying "false positives" in information seeking, i.e. user satisfaction with poor search results. He concludes: There appears to be little interaction between these two variables [actual search performance, i.e. quality and relevance of results, and user satisfaction]. Searchers may express satisfaction with search results even when the results are far from optimal."

[8]The Web Archive "Wayback Machine" http://web.archive.org launched for public access in October 2001

Studies of typical user's search behavior and preferences have strongly influenced the evolution of online searching and suggest what kinds of search engines will thrive in the years to come.[9] In general, users show a marked preference for directories with pre-selected links organized by topic or for sites with a natural-language interface such as AskJeeves over full-text search engines like AltaVista or Google. At sites like the latter, 80%-90% of all queries consist of a single word or phrase. While AltaVista supports complex queries with Boolean operators (logical operators like AND, OR, NEAR, NOT) and bracketing, up to 25% of such queries submitted are ill-formed and thus return no results. Users tend to follow up only the first few hits in the search results, calling them up one for one in the same window, then returning to get the next link.

Extrapolating from such studies and from current trends in user figures, it appears that "geek seek" full-text search sites like AltaVista will decline to the benefit of less powerful search engines which offer cleaner, more accommodating user interfaces and higher ranking for the results with the greatest likely relevance. Unfortunately for language professionals, it is precisely the complex queries rooted in the arcane world of Unix and grep that facilitate targeted online linguistic research. AltaVista's successful challenger Google[10] has prospered because its link popularity ranking usually yields relevant results—and because its coverage of the Web is vast and up to date. While it is a full-text search engine, its support for Booleans is limited to AND, OR and NOT; it lacks NEAR, wildcards and bracketing, and its distinction between lower and upper case and between plain characters and those with diacritics is inconsistent. Worse yet, Google's link popularity ranking works against diversity in the search results. Perhaps the most unsettling trend for linguistic investigation is the development of information retrieval search models and natural language user interfaces. While a boon for novice searchers (and NLP researchers), these approaches will favor the largest languages from the wealthiest countries, excluding those for which linguistic data are already most difficult to obtain.

Genesis and Development of KWiCFinder

After the launch of AltaVista in 1995 I became an intensive search-engine user. Soon I learned how to maximize online search efficiency despite a slow connection: I would get a page of hits, open each in a new window, go back to the search engine for more hits, then evaluate the pages that had loaded in the meantime. Occasionally I would go off leaving a couple of dozen documents to load in my absence for subsequent perusal. Only a small percentage of students and colleagues to whom I tried to teach my multitasking method adopted this approach; the rest continued to express frustration with the large amount of time spent sifting through hits to find relevant webpages. It occurred to me that I could automate the process of search and retrieval by writing a program to submit the query to AltaVista, then retrieve the pages and save them to disk automatically. This first step, dubbed WebFetch, satisfied my own immediate needs, but had little appeal for my students, since they still had to open and peruse numerous downloaded files. To expedite evaluation of those webpages, I started excerpting the webpages and producing reports with KWiC display, resulting in KWiCFind, which several volunteers from my students evaluated in the spring of 1997. When our institution finally left Windows 3.1 behind, I

---

preserves 10 billion webpages that have changed or vanished.

[9]Relevant user studies include Körber 2000; Jansen, Spink and Saracevic 2000; Silverstein, Henzinger, Marais and Moricz 1999.

[10]The author's webpage http://miniappolis.com/KWiCFinderVSWebKWiC.html details the primary differences between Google and AltaVista.

reprogrammed it from the ground up for 32-bit Windows, providing the specific enhancements for foreign language users and linguists detailed below. At the 1999 CALICO Conference the reborn KWiCFinder was shown for the first time outside my classroom (Fletcher 1999). The current version of KWiCFinder can be downloaded free of charge at http://miniappolis.com/KWiCFinder/.

AltaVista offers a combination of features that make it the most powerful search engine to support. Unlike many others, it indexes **all** words, including the frequent "stopwords" ignored by others which may be the focus of linguistic investigation. To a certain extent it allows queries which distinguish upper-case letters from lower-case ones and "special" characters with diacritics from their "plain" counterparts. It even has some language-specific knowledge, for example about the equivalence of *ä* and *ae, ß* and *ss* in German. It provides true world-wide coverage and was the first to offer search by language. Essential for targeted searches, it supports Boolean operators, bracketing, and wildcards, and imposes no limits on the length or complexity of a query. Finally, AltaVista performs literal text matching, without attempting to "second guess" the user's intent. After having been sold and reorganized several times, AltaVista's market share has diminished significantly, especially in the USA. It lags far behind some rivals in size and freshness of its content, and stands out with the highest percentage of dead links among the major search engines.[11] Nonetheless its support for complex queries still makes it a very useful tool.

Daily experience with KWiCFinder and frustration with search engines led me to refine wildcard matching strategies to reduce false matches. "Wildcards" permit a search term to match likely variants of a given word without the user's entering each alternate form. For example, the AltaVista wildcard symbol * matches any sequence of zero to five characters, so the search term *nation** would match singular / plural forms like *nation, nations,* as well as derived words like *national, nationalism, nationality, nationalize/ise* etc., and *labo*r* matches both American *labor* and British *labour.* Furthermore, AltaVista 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 would match any of *aáâäàãæåAÁÂÄÀÃÆÅ).* These "implicit wildcards" ensure that many paradigmatic and graphic variants of a given word match a single search term, despite the differences introduced by factors like sentence-initial capitalization; required, omitted or misused diacritics; or alternate spellings due to keyboard limitations.

While wildcards increase the efficiency of *entering* search terms, they can also lead to many irrelevant matches which must be sifted out individually. To address this problem I implemented single-character wildcards and the "sic" option in KWiCFinder. Borrowing from standard concordance practice, I added the wildcard characters *?* and *%* to the inventory to match either one (no more, no less) or zero to one character respectively. KWiCFinder's "sic" option forces a plain or lower-case character in a search term to match only that exact character. Similarly, to AltaVista's native NEAR Boolean operator, which requires only that one search term be within ten words on either side of another term, I added BEFORE and AFTER operators, and permitted users to specify a shorter distance between the terms. All these enhancements reduce the likelihood of unwanted matches.

These refinements—single-character wildcard and "sic" matching as well as specifying relative order and degree of proximity of two terms—do come at a significant price. When a query is submitted, it can only be as specific as the search engine's conventions allow. A more general search may match many documents which must be discarded after retrieval and analysis because they do not actually meet the user's more specific criteria. In one intentionally extreme test I invoked "sic" to seek examples of the

[11]http://www.searchenginewatch.com tracks numerous search engine developments and statistics.

German verb form *(ihr) fahrt* 'you (plural) go', far less frequent than either the special-character third-person singular *fährt* or the capitalized noun *Fahrt.* KWiCFinder had to retrieve over 200 documents matching the search term *fahrt* according to AltaVista's criteria to find a single citation of the desired form![12] However, since the program automates the entire process, even in such an extreme case it does use *human* time very efficiently.

The most efficient searches result from queries which avoid wildcards and specify every alternate search term completely. Nevertheless, entering all desired variants of a given form can be daunting and highly repetitive, especially in languages with richer morphology than English. To transfer this tedious task to the machine, I introduced "tamecards", a shorthand for generating alternate forms. For example, KWiCFinder expands the tamecard notation *s[iau]ng[,s,ing]* to all forms of the verb *sing: sing, sings, singing, sang, sung* (as well as the nonsense forms *sangs, sungs, sanging, sunging,* which fortunately yield no false matches). Each of these forms is then submitted to the search engine so that only perfect matches are retrieved. Since derivational and inflectional patterns typically apply to many words, such tamecard formulas can be saved, then pasted in as needed. A further refinement is the "indexed tamecard", in which every *n*th field in curly braces corresponds to the corresponding field in other sets of curly braces within the same search term, so that *{me,te,se} lav{o,as,a}* expands to *me lavo, te lavas, se lava.* Such shorthand for fully-specified alternate forms would be a boon to searching on sites which do not support wildcards such as Google.

Another pair of KWiCFinder tamecard conventions addresses orthographic inconsistency in compounds which can be written as one word or two, either joined by a hyphen or separated by a space. A hyphen or apostrophe in a search term is expanded to alternate forms with or without a space.[13] Consequently, *on-line* matches any of the interchangeable 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 contemporary German (as is AltaVista's lower-case / upper-case equivalence), which now is in a ten-year period of transition to a new spelling. The reforms permanently separate many words formerly written as one, while fusing some former phrases into single words; they also allow individual discretion in breaking up German's notoriously long compounds with hyphens, leading to even greater orthographical variation. While the media and most schools are implementing the new spelling, many online sources will continue to reflect traditional orthography for years. With KWiCFinder, the search term *kennen-lernen* matches both old-style *kennenlernen* and reformist *kennen lernen.* This tamecard convention provides a simple means of matching both variants with a single entry.

In addition to these enhancements to query formulation KWiCFinder introduces a further means of narrowing a search, "inclusion" and "exclusion" criteria. These may be words whose appearance on a webpage helps target a specific domain or, alternatively, disqualifies that page from further consideration; these terms are submitted to the search engine as part of the query, but do not appear in KWiCFinder's search 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. As exclusion criteria these latter parameters help one filter out unwanted material.

Once launched, KWiCFinder works without further attention, retrieving five to ten documents a minute, excerpting them, and finally producing a search report which displays the key search terms in the amount

[12]Searching for *ihr fahrt* OR *fahrt ihr* would be the efficient way to do this.

[13]Like other search engines, AltaVista treats punctuation marks as spaces.

of context specified by the user, along with information on and links to the source documents. Multiple independent searches can be carried out simultaneously, which is especially beneficial for long unattended searches. To expedite later review, one can choose to save the original documents on the hard drive in HTML and / or text format. These original texts are then instantly available offline for perusal, editing and reproduction, or for further analysis by a full-featured concordancing program, and they remain accessible even if the online version is changed or removed.

KWiCFinder's user report options have always offered various ways to set off the keywords from the surrounding text, and allowed a choice between a single report document per search and individual reports per document, practical for rapidly evaluating documents in a more extensive search. Recent stabilization of the XML (eXtensible Markup Language) encoding and XSLT (eXtensible Stylesheet Language Transformation) rendering standards have permitted KWiCFinder to offer an additional highly versatile report format since mid-2000.

XML provides a standard method for tagging structured data in a text file format that can be easily understood by both humans and computers. While HTML offers the page designer (in this case the KWiCFinder programmer) reasonable control over page *appearance,* its formatting markup tags furnish no clues to the *structure* of the information on the page; once an HTML page has been completed, its form is basically set. In contrast, XML has no built-in display formatting, but provides a standard approach to defining and encoding the structure of the information, essentially as a user-defined database. Consider this simplified snippet of a citation from an XML-encoded search report. Programmer-defined tags identify components as "<precontext>", "<matchingtext>", or "<postcontext>".

```
<cite citeID="8.1.1">
<precontext>
   Da das LRZ anfangs mit ähnlichen Gerätschaften zu tun hatte
   der erste Rechner hieß PERM, natürlich nicht nach dem Erdzeitalter,
</precontext>
<matchingtext>
   sondern
</matchingtext>
<postcontext>
   als Abkürzung für "programmgesteuerte elektronische Rechenanlage
   München"- könnte man hier den ersten Zusammenhang sehen.
</postcontext>
</cite>
```

All of the data from a KWiCFinder search are stored in this way in an XML file. To generate a useful report, KWiCFinder applies an XSLT "stylesheet" to this database to select which information to display, insert appropriate text labels in the desired language, and format the result as an HTML document for display in its browser window.

The advantages and power of XML encoding becomes clear from the samples of an actual search report accessible via this link. Display *form* is perfectly separated from search report *content,* and it can be modified as needed. To change the display format or the language of the text labels, KWiCFinder merely applies a different stylesheet to the same XML file; there is no need to reanalyze the original documents. With appropriate knowledge of XSLT and browser scripting techniques, an end user could create new report formats or apply other stylesheets to annotate, merge, prune, or restructure XML search reports. There are numerous instructive examples of these manipulations online (at sites like http://www.xml.org, http://www.xml.com and http://msdn.microsoft.com/xml/) and in books, such as Britt and Duynstee

(2000); Kay (2000) provides a comprehensive reference to XSLT. While learning to work with these technologies is not a trivial enterprise, the growing commercial enthusiasm for XML promises that this expertise will continue to become more readily available. The ability to perform sophisticated database and report display manipulations in a current-generation browser points the way to a future cross-platform approach to learner concordancing.

WebKWiC

Some searchers have been intimidated by the effort required to download, install, and learn to use KWiCFinder, yet they still can benefit from automation of search and retrieval. To lower the entry threshold for such users I created WebKWiC,[14] a light-weight, fully browser-based JavaScript application. It capitalizes on Google's "Document from Cache" feature, which serves up a copy of a webpage matching a user's query from Google's archives, highlighting instances of the search terms with color codes. WebKWiC retrieves several of these cached pages at a time and adds buttons so the user can navigate easily among citations and windows, greatly enhancing the efficiency of previewing large numbers of documents. WebKWiC also adds a means of entering "special characters" to the user interface and gives certain essential search options greater prominence than does Google's original page. Google is an ideal partner for an entry-level search agent like WebKWiC. Its straightforward approach to advanced search with "implicit Booleans" is easy to learn, so users either come equipped with or acquire readily transferrable skills. Since Google indexes major non-Western European / non-Roman orthography languages, this approach allows me to meet the needs of a population which KWiCFinder does not support yet.[15]

Webidence as Evidence

We all know (and may ourselves have voiced) the complaints about online information: there is too much ephemeral content of dubious reliability; journalistic, commercial and personal texts of unknown authorship and authority abound; assertions are intermingled with and represented as established fact, and details of sources and research methodology are documented haphazardly at best. For linguistic research even more caution is essential for numerous reasons. The Internet domains in a URL (e.g. .ca, .uk, .de, .jp, .com, .edu) are only a rough guide to provenance. In addition, many webpages consist primarily of fragments–titles and captions, supplemented by the occasional imperative ("click here for more information", "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, at the same time fusing creolized Web English into texts in their native tongue. Similarly, while searching

[14]http://miniappolis.com/WebKWiC/

[15]Recently a couple of alternatives to KWiCFinder and WebKWiC have appeared. Two online pages produce KWiC concordances from search results: WebCORP (http://www.webcorp.org.uk), which uses various search engines and provides a number of analytical tools, and WebCONC (http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi), which works with Google only. Both offer a distinct advantage: processing takes place on the server, so no software needs to be downloaded, and users with slow connections can concordance large numbers of documents in a relatively short time; neither has the search-engine extensions, language enhancements or reporting flexibility of KWiCFinder. A third possibility, TextSTAT, (http://www.niederlandistik.fu-berlin.de/taalkunde/software.html), will retrieve and create a simple concordance of a URL entered by the user. Finally, the search agent Copernic (http://www.copernic.com, available in a free "basic" version) performs searches on multiple sites simultaneously and shows one instance from each document of a search term in a very brief context.

for linguistic examples I often have stumbled upon compositions by learners with imperfect mastery of the language (many language courses post student work for peer review) as well as numerous baffling documents that turned out to be machine translated.[16] In many online discussion groups, sloppy spelling and careless language appear to be the norm. With its frenetic pace of development, the Web typically values content *creation* above content *perfection* and tolerates ill-formed language–after all, those who are upset by it can find relief a click away.

In light of these pitfalls our profession needs to develop "Standards of Webidence" to guide the selection and documentation of online language for linguistic research. We also must understand and beware of the limitations of search engines. In particular, the number of pages matching a query reported by a search engine gives a rough numerical indication at best; comparison of these numbers does not constitute statistical proof.[17] Search engines report **the number of pages** matching a query, **not the actual number of citations** on those pages. A single page may contain several alternate usages (as in the *los sesentas* example above), thus appearing in more than one count. On the other hand, numerous pages may propagate verbatim a formulation originating in a single document, thus multiplying its frequency, as in copied quotations, song lyrics, aphorisms, anecdotes, and jokes; online forums in which an original posting and all subsequent comments are repeated in each successive posting; and mirror sites for FADs (frequently-accessed documents). Furthermore, a single site may be responsible for most or all the hits of a query for a spurious or unusual usage.

AltaVista itself warns not to trust its figures: when its servers experience heavy traffic, generating result pages receives priority over producing hit counts, so numbers for the same query easily vary by an order of magnitude over the course of one search session.[18] Finally, the fact that a given form or construction can be found on the Web does not amount to proof of its existence in a language: many hapless hapax legomena born of input error or syllable stranding by hyphenation wait on the Web for an unsuspecting searcher to united them with their orphaned siblings.

KWiCFinder facilitates responsible online linguistic scholarship in several ways. It allows one to review large numbers of documents and citations efficiently, with each keyword shown in sufficient context to evaluate its relevance and validity. It can tally the number of instances of each keyword in a document for calculation of its relative frequency. The user can choose to save the documents to a local file to permit further analysis or independent verification of results. It incorporates tools to annotate, classify and delete individual citations or entire documents from a search report. Finally, complementary corpus analysis tools now under development enable one to eliminate unrepresentative, redundant or repetitive documents from further consideration.

---

[16]One machine translation was so artless that even the HTML tags were rendered in Spanish, with <*CABEZA*> and <*CUERPO*> replacing <*HEAD*> and <*BODY*>*!*

[17] It is unclear how many linguists are aware of these limitations. I have seen postings in scholarly fora such as Linguist List citing hit counts from AltaVista as evidence prevalence of a given form over another with no indication that the poster either has followed up to verify a substantial number of the hits or is even aware of the limitations of this method.

[18] The page "AltaVista Advanced Search Tutorial--About the Page Count" cited in the bibliography explains this limitation of AltaVista's hit counts. Brekke (2000) and Meyers et al. (2001) note this problem as well. I have even received *negative* hits counts like "We found -40,000,000 results." To ensure the most accurate counts, follow AltaVista's advice by accessing it at off-peak times, e.g. on weekend mornings. Note that specifying "one page per site" does not affect total counts, so it provides no indication of how widespread a given form or construction is.

REFERENCES

Ackermann, E., & Hartman, K. (2000). *The Information Specialist's Guide to Searching and Researching on the Internet & the World Wide Web.* Wilsonville, OR: ABF Content.

Agence de la Francophonie. (2001). *The Fifth Study on Languages and the Internet.* Retrieved 18 October 2001 from the World Wide Web: http://funredes.org/LC/english/L5/L5contents.html

AltaVista Advanced Search Tutorial - About the Page Count. Palo Alto, CA: AltaVista Company. Verified 11 November 2001 on the World Wide Web: http://help.altavista.com/adv_search/ast_as_pagecount

Aston, G. (2001). *Learning with Corpora.* Houston: Athelstan.

Banko, Michele and Brill, Eric. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. ACL 2001. Retrieved 5 October 2001 from the World Wide Web: http://research.microsoft.com/~brill/Pubs/ACL2001.pdf

Bernardini, Silvia. (2000). Systematising Serendipity: Proposals for Concordancing Large Corpora with Language Learners. In Burnard and McEnery, 224-34.

Bernardini, Silvia. (2000). Serendipity expanded: Exploring new directions for discovery learning. Fourth International Conference on Teaching and Language Corpora, Graz. Abstract retrieved 26 July 2001 from the World Wide Web: http://www-gewi.kfunigraz.ac.at/talc2000/Dokumente/abstracts/Bernardini%20Abstract.doc

Bickel, H. (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. (pp. 111-124)* Tübingen: Francke. Retrieved 12 October 2001 from the World Wide Web: http://www.germa.unibas.ch/seminar/whoiswho/Publikationen/Variationsling.pdf

Brekke, M. (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) (pp. 227-247).* Amsterdam, Atlanta: Rodopi.

Britt, J. & Duynstee, T. (2000). *Professional Visual Basic 6 XML.* Birmingham UK: Wrox Press.

Burnard, Lou and Tony McEnery (eds) (2000). Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora. Frankfurt am Main: Peter Lang. (□od□ Studies in Language Vol 2).

Cavaglià, Gabriela and Adam Kilgarriff. (2001). "Corpora from the Web". Fourth Annual CLUCK Colloquium, Sheffield, UK, January 2001. Retrieved 8 November 2001 from the World Wide Web: ftp://ftp.itri.bton.ac.uk/reports/ITRI-01-11.pdf

Grefenstette, Gregory. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. ASLIB, Translating and the Computer 21, London, Nov 10-11, 1999. Retrieved 12 November 2001 from the World Wide Web: http://www.xrce.xerox.com/research/mltt/publications/Documents/P49030/content/gg_aslib.pdf

Hildreth, Charles R. (2001). Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study. Information Research, 6/2. Retrieved 22 January 2001 from the World Wide Web: http://InformationR.net/ir/paper101.html
Hilton, James. (2001). Copyright Assumptions and Challenges. EDUCAUSE Review 36/6, November / December, 48-55. Retrieved 12 November 2001 from the World Wide Web: http://www.educause.edu/ir/library/pdf/erm0163.pdf

Crews, Kenneth D. (2000). Fair Use: Overview and Meaning for Higher Education. Retrieved 12 November 2001 from the World Wide Web: http://www.iupui.edu/~copyinfo/highered2000.html

Estadísticas de Internet en el ámbito internacional Madrid: Asociación de Usuarios de Internet. Retrieved 6 November 2001 from the World Wide Web: http://www.aui.es/estadi/internacional/internacional.htm

Evolution & Projections of Online Linguistic Populations. San Francisco, CA: Global Reach Retrieved 6 October 2000 from the World Wide Web: http://www.glreach.com/globstats/evol.html

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

Franz, Alex. 2001. Cite Google.com e-mail Nov

Grefenstette, Gregory & Julien Nioche. (2000) Estimation of English and non-English Language Use on the WWW. RIAO 2000, Paris, 12-14 April 2000. Retrieved 12 October 2001 from the World Wide Web: http://www.xrce.xerox.com/research/mltt/publications/Documents/P19137/content/RIAO2000gref.pdf

Global Internet Statistics (by Language). San Francisco, CA: Global Reach Retrieved 6 October 2001 from the World Wide Web: http://www.glreach.com/globstats/index.php3

Godwin-Jones, Robert. (1999). Web Metadata: More Efficient Resource Cataloging and Retrieving. *Language Learning & Technology* 3, (1): 12-16

Grefenstette, Gregory (1999) The World Wide Web as a Resource for Example-Based Machine Translation Tasks. Retrieved 12 October 2001 from the World Wide Web: http://www.xrce.xerox.com/research/mltt/publications/Documents/P49030/content/gg_aslib.pdf

Grefenstette, Gregory & Julien Nioche. (2000) Estimation of English and non-English Language Use on the WWW. RIAO 2000, Paris, 12-14 April 2000. Retrieved 12 October 2001 from the World Wide Web: http://www.xrce.xerox.com/research/mltt/publications/Documents/P19137/content/RIAO2000gref.pdf

Halteren, Hans van (1999). Tekstcorpora en taalgerelateerd onderwijs. *STDH Nieuwsbrief, 9* (Feb.): 7-9.

Hawkins, Donald T. (1996). Hunting, Grazing, Browsing: A Model for Online Information Retrieval.

ONLINE 20(1 January). Retrieved 21 October 2001 from the World Wide Web: http://www.onlinemag.net/JanOL/hawkins.html

Hock, Randolph. (1999). *The Extreme Searcher's Guide to Web Search Engines. A Handbook for the Serious Searcher.* Medford, NJ: Information Today. Supplement and update at http://www.onstrat.com/engines/

Jansen, B. J. (2000) The Effect of Query Complexity on Web Searching Results. *Information Research, 6*(1). Verified 16 November 2001 on the World Wide Web: http://informationr.net/ir/6-1/paper87.html

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, *36* (2): 207-227.

Jansen, B. J. & Pooch, U. (2001) A Review of Web Searching Studies and a Framework for Future Research. *Journal of the American Society for Information Science and Technology, 52* (3): 235-246.

Johns, Tim F. (2000). Tim Johns English for Academic Purposes Page. Retrieved 18 October 2001 from the World Wide Web: http://web.bham.ac.uk/johnstf/timeap3.htm#revision

Johns, Tim F. (2001). Modifying the Paradigm. Third North American Symposium on Corpus Linguistics and Language Teaching, Boston, MA, 23-25 March 2001.

Kilgarriff, Adam (2001). "Web as corpus". In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja (eds.) *Proceedings of the Corpus Linguistics 2001 conference, UCREL Technical Papers: 13.* Lancaster University, 342-344.

Kay, Michael. (2000). *XSLT Programmer's Reference.* Birmingham UK:Wrox Press.

Körber, Sven. (2000). Suchmuster erfahrener und unerfahrener Suchmaschinennutzer im deutschsprachigen World Wide Web. Ein Experiment. Unpublished master''s thesis, Westfälische Wilhelms-Universität Münster, Germany. Retrieved 14 September 2001 from the World Wide Web: http://kommunix.uni-muenster.de/IfK/examen/koerber/suchmuster.pdf

Lamy, M. N., Klarskov Mortensen, H. J. & Davies, G. (2000). Using concordance programs in the modern foreign languages classroom. ICT4LT Module 2.4. Retrieved 1 June 2001 from the World Wide Web: http://www.ict4lt.org/en/en_mod2-4.htm

Lawrence, S. & C. L. Giles. (1998). Context and Page Analysis for Improved Web Search. *IEEE Internet Computing, 2* (4):38-46.

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

Limberg, L. (1999) Experiencing information seeking and learning: a study of the interaction between two phenomena. *Information Research 5*(1). Retrieved 17 September 2000 from the World Wide Web: http://www.shef.ac.uk/~is/publications/infres/paper68.html

Meyer, Charles Roger Grabowski, Thomas Han, Konstantin Mantzouranis, & Stephanie Moses. (2001). The World Wide Web as Linguistic Corpus. Third North American Symposium on Corpus Linguistics

and Language Teaching, Boston, MA, 23-25 March 2001.

McEnery, T. & Wilson A. (1996). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McEnery, T. & Wilson, A. (2000). Corpus linguistics. ICT4LT Module 3.4. Retrieved 1 June 2001 from the World Wide Web: http://www.ict4lt.org/en/en_mod3-4.htm

Moore, A. & Murray, B.H. (2000). Sizing the Internet. July 10, 2000. Arlington, VA: Cyveillance, Inc. Retrieved 8 October 2000 from the World Wide Web: http://www.cyveillance.com/resources/7921S_Sizing_the_Internet.pdf

Nua Internet How Many Online. Dublin: Nua Ltd. Retrieved 8 October 2001 from the World Wide Web: http://www.nua.ie/surveys/how_many_online/index.html and regional subpages.

Pearson, Jennifer (2000). Surfing the Internet: teaching students to choose their texts wisely. In Burnard and McEnery, 235-39.

Petitclerc, Angela (1998). Corpora for Concordances. Paper for course assignment. Montreal: Concordia University. Retrieved 12 March 2000 from the World Wide Web: http://rkenner.concordia.ca/GSE555_98/A_Petitclerc/corpora.html

Reibold, H. & Neumeier, F. (2000) Findigkeit gefragt. *PC Professionell* (February) 177-ff. Retrieved *7* July 2000 from the World Wide Web: http://www.zdnet.de/internet/artikel/scene/200002/suchmaschinen01_00-wc.html

Salzmann, A. & Mills, D. (2000). LinguaCenter Grammar Safari. Retrieved 23 July 2000 from the World Wide Web: http://deil.lang.uiuc.edu/resources/web.pages/grammarsafari.html

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33(1), 6 -12. Retrieved 11 October 2000 from the World Wide Web: http://www.acm.org/sigir/forum/F99/Silverstein.pdf

Size of the web: A Dynamic Essay For a Dynamic Medium Retrieved 1 October 2000 from the World Wide Web: http://censorware.org/web_size/

Spink, A. & Xu, J. L. (2000). Selected Results from a Large Study of Web Searching: the Excite Study. *Information Research, 6*(1). Retrieved 16 October 2000 from the World Wide Web: http://www.shef.ac.uk/~is/infres/paper90.html

Volk, Martin (2000). Scaling up. Using the WWW to resolve PP attachment ambiguities. *Proceedings of Konvens-2000, Sprachkommunikation,* Ilmenau, VDE Verlag, 151-156. Retrieved 12 October 2001 from the World Wide Web: http://www.ifi.unizh.ch/cl/volk/papers/Konvens2000_Ilmenau.pdf

Volk, Martin (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. Retrieved 12 October 2001 from the World Wide Web: http://www.ifi.unizh.ch/cl/volk/papers/Lancaster_2001.pdf

Whistle, J. (1999). Concordancing with Students Using an ‘‘Off-theWeb’’ Corpus. *ReCALL 11* (2):74-80.

Zanettin, Federico. (2001a) Swimming in Words: Corpora, Translation and Language Learning. In Aston 177-197.

Zanettin, Federico. (2001b) DIY Corpora: the WWW and the Translator. Training the Language Services Provider for the New Millennium, Porto, Portugal, 25-26 May 2001. Retrieved 22 January 2002 from the World Wide Web: http://www.federicozanettin.net/DIYcorpora.htm